Journal *of* Accounting Research

DOI: 10.1111/j.1475-679X.2012.00450.x
Journal of Accounting Research
Vol. 50 No. 2 May 2012
Printed in U.S.A.

CHICAGO BOOTH

Detecting Deceptive Discussions in Conference Calls

DAVID F. LARCKER* AND ANASTASIA A. ZAKOLYUKINA†

Received 23 July 2010; accepted 23 December 2011

ABSTRACT

We estimate linguistic-based classification models of deceptive discussions during quarterly earnings conference calls. Using data on subsequent financial restatements and a set of criteria to identify severity of accounting problems, we label each call as "truthful" or "deceptive." Prediction models are then developed with the word categories that have been shown by previous psychological and linguistic research to be related to deception. We find that the *out-of-sample* performance of models based on CEO and/or CFO narratives is significantly better than a random guess by 6–16% and is at least equivalent to models based on financial and accounting variables. The language of deceptive executives exhibits more references to general knowledge, fewer nonextreme positive emotions, and fewer references to shareholder value. In addition, deceptive CEOs use significantly more extreme positive emotion and fewer anxiety words. Finally, a portfolio formed from firms with the highest deception scores from CFO narratives produces an annualized alpha of between −4% and −11%.

*Graduate School of Business, Rock Center for Corporate Governance, Stanford University, †Graduate School of Business, Stanford University. We would like to thank Thomas Quinn for his help in securing the FactSet data, and John Johnson and Ravi Pillai for their help with computational issues. The comments of Maria Correia, Jerome Friedman, Michelle Gutman, Daniel Jurafsky, Sally Larcker, Andrew Leone, Sergey Lobanov, Miguel Angel Minutti Meza, Maria Ogneva, Brian Tayan and participants at the Transatlantic Doctoral Conference 2010 at the London Business School, American Accounting Association Meeting 2010, and 2011 Journal of Accounting Research Conference for helpful discussions. We thank Philip Berger (the Editor), the anonymous referee, and Robert Bloomfield for their excellent suggestions. Larcker also thanks the Joseph and Laurie Lacob Faculty Fellowship for financial support.

Copyright ©, University of Chicago on behalf of the Accounting Research Center, 2012

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. [2011]). In addition, professional organizations, such as Audit Integrity Inc. (now GovernanceMetrics International), have developed commercial models that claim to provide warning signs of managerial manipulation of accounting reports. Despite extensive prior research, the ability of these models to identify and predict accounting manipulations is quite modest.

In this paper, we take a different approach to the prediction of financial statement manipulations. Rather than use heuristics based on accounting relations, we analyze linguistic features present in CEO and CFO statements during quarterly earnings conference calls. In particular, we examine the formal Management Discussion (MD) and Question and Answer (Q&A) narratives in conference calls for linguistic features that predict "deceptive" reporting of financial statements. Our study is based on the considerable prior work in linguistics, psychology, and deception detection research, which finds that the language composition of truthful narratives differs from that of false narratives. Our primary assumptions are that CEOs and CFOs know whether financial statements have been manipulated and formal and spontaneous narratives of these executives provide cues that can be used to identify deceitful (or lying) behavior.

Using a comprehensive set of electronic transcripts for quarterly conference collected by FactSet Research Systems Inc. and restatements identified by Glass, Lewis and Co., we build prediction models for the likelihood of deception during the September 2003 to May 2007 time period. Four different alternative methods are used to label a conference call narrative as "deceptive." The first approach labels a restatement as deceptive if it involves one of the following: a disclosure of a material weakness, an auditor change, a late filing, or a Form 8-K filing. The second approach labels a restatement as deceptive if it relates to an irregularity as described in Hennes, Leone, and Miller [2008] or if it involves accounting issues that elicit a significant negative market reaction such as those described in Palmrose, Richardson, and Scholz [2004] and Scholz [2008]. The third approach labels a restatement as deceptive if it involves an irregularity as defined in Hennes, Leone, and Miller [2008]. Finally, the fourth approach labels a restatement as deceptive if the restatement involves a formal SEC investigation that leads to an issuance of an Accounting and Auditing Enforcement Release (or AAER).

In *out-of-sample* tests, our linguistic classification models based on CFO (CEO) narratives perform significantly better than a random guess by 6–16%. We also find that the models based on linguistic categories have statistically better or equivalent predictive performance compared to various accounting models that rely on discretionary accruals and the commercial accounting score developed by Audit Integrity Inc.

In terms of the linguistic features, deceptive CEOs and CFOs use more references to general knowledge, fewer nonextreme positive emotion words, and fewer references to shareholder value. We also find substantial differences between CEOs and CFOs. Deceptive CEOs use significantly more extremely positive emotion words and fewer anxiety words. In contrast, deceptive CFOs do not use extremely positive emotion words. However, they use significantly more words of negation and extremely negative emotion words. These results are generally consistent with prior theoretical and empirical studies of deception in psychology and linguistics.

Finally, to assess the economic relevance of our linguistic models, we construct monthly calendar time portfolios of firms according to deception scores of their conference calls. Each firm is selected to the equally weighted portfolio for three months. The composition of the portfolios is updated monthly. The annualized alpha (estimated using the four-factor model) for this portfolio selection strategy using the CFO linguistic model is between a negative 4% and 11%, depending on the deception criterion and score percentile. The results for the CEO linguistic models do not produce a statistically significant alpha.

Overall, our results suggest that the linguistic features of CEOs and CFOs in conference call narratives can be used to identify financial misreporting. 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 that high levels of classification performance are unlikely for this initial study. Despite this caveat, we believe that our initial exploratory results suggest that it is worthwhile for researchers to consider linguistic features when attempting to measure the quality of reported financial statements.

The remainder of the paper consists of six sections. Section 2 provides a review of prior accounting and finance work analyzing the linguistic features of press releases, formal Securities and Exchange Commission (SEC) filings, and other similar text documents. Section 3 discusses the theoretical background used to justify our choice of word categories. The sample construction is discussed in section 4, and measurement and econometric choices are developed in sections 4 and 5. Our primary results for the linguistic prediction models are presented in section 6. Concluding remarks, limitations, and suggestions for future research are provided in section 7.

2. *Prior Research Analyzing Linguistic Features*

Several recent papers in accounting and finance have analyzed various linguistic features in formal corporate disclosures (e.g., Demers and Vega [2010], Li [2006, 2008, 2010], Loughran, McDonald, and Yun [2009]), press releases (e.g., Davis, Piger, and Sedor [2007], Henry and Leone [2009]), media news (e.g., Tetlock [2007], Tetlock, Saar-Tsechansky, and Macskassy [2008], Core, Guay, and Larcker [2008]), and Internet message boards (e.g., Antweiler and Frank [2004], Das and Chen [2007]). Many of these studies differ in terms of the linguistic cues under consideration and the techniques used to extract these features. For example, some studies count the frequency of particular words, whereas others analyze overall positive (optimistic) or negative (pessimistic) tone in the text. Researchers have used hand-collected lists of words, simple word counts from psychosocial dictionaries, and estimates produced by various 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 the risk sentiment of annual 10-K filings where risk sentiment is measured by counting words related to risk ("risk," "risks," and "risky") and uncertainty ("uncertain," "uncertainty," and "uncertainties"). Core, Guay, and Larcker [2008] analyze newspaper articles about CEO compensation and identify articles that have negative tone by keywords. Similarly, Loughran, McDonald, and Yun [2009] collect and analyze a list of ethics-related terms in 10-K annual reports. However, hand-collected word lists can be confounded by researcher subjectivity and miss important dimensions that may be captured by more comprehensive psychosocial dictionaries and automatic classifiers. Conversely, an important advantage of hand collection is that the researcher must identify the linguistic constructs of interest and the precise words that are related to these constructs.

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, Li, and Short [2009] also use GI to count negative and positive words in disclosures by management, analyst reports, and business news. Davis, Piger, and Sedor [2007] measure linguistic style (tone) in press releases for earnings using the software package named Diction. The study measures words that are optimistic (e.g., praise, satisfaction, and inspiration) or pessimistic (e.g., blame, hardship, and denial) as a percentage of all words in press releases.

Perhaps the most important criticism associated with using word counting ("bag-of-words" approach) based on psychosocial dictionaries is that this approach does not differentiate between several meanings of words with the same appearance in the text. Pure word counting also does not categorize combinations of words (or phrases) that might imply different meanings from the constituent words. Another important issue is that most of the general dictionaries are not compiled for analyzing business communication. This raises questions about whether the dictionary contains the necessary set of words for business communication. Despite these limitations, the "bag-of-words" approach is simple, parsimonious, and replicable.

Another methodological approach is to apply text classifiers (such as the common Naive Bayesian algorithm) from computational linguistics. 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, or sell categories. Similarly, Balakrishnan, Qiu, and Srinivasan [2010] use text classification to assign manufacturing firms as out-/underperforming based on narrative disclosures in their 10-K filings. Li [2010] concludes that the tone measures estimated by the Naive Bayesian classifier applied to forward-looking statements in the MD&A section of 10-K and 10-Q filings exhibit a statistically positive association with future performance, whereas the tone measures extracted using traditional dictionaries (Diction, GI, and LIWC) are not associated with future performance.

Some prior studies explicitly examine the readability and related obfuscation of written disclosures such as prospectuses or 10-K reports (e.g., Courtis [2004], Li [2008], Humpherys et al. [2011], Loughran and McDonald [2011]). For instance, Li [2008] examines annual report disclosures by counting linguistic features related to obfuscation, such as the relative frequency of self-reference words, causation words, positive emotional words, and future tense verbs. He finds that more causation words, less positive words, and more future tense verbs are associated with obfuscation as measured by less persistent positive earnings. At the same time, Loughran and McDonald [2011] find that negative, uncertainty, and litigious words in 10-Ks are statistically significant in predicting 10b-5 fraud lawsuits only when the word categories are weighted to account for the words rarity.

In contrast to prior studies on deception that use the text in written disclosures such as 10-Ks, we use arguably more spontaneous disclosures of conference calls. There are a number of limitations to using formal disclosures such as 10-Ks and 10-Qs in deception studies. First, formal disclosures are more scripted and prior research has shown that their content does not change much over time. Second, the different parts of the reports are written and edited by different individuals and these individuals are unlikely to be executives. Finally, these disclosures lack the spontaneity that characterizes conference calls.

Prior accounting and finance research has uncovered a number of provocative associations between linguistic cues and firm performance. However, with the possible exception of obfuscation analysis of written disclosures by Li [2008], Loughran and McDonald [2011], and Humpherys et al. [2011], there is little prior work using linguistic features to identify deceptive reporting behavior by corporate executives. The purpose of this paper is to use contemporary linguistic methods and analysis to develop a predictive model for deception (or lying) by CEOs and CFOs during quarterly conference calls.

3. Development of Word Categories

3.1 THEORETICAL BACKGROUND

We select our word categories based on the extensive review and synthesis provided by Vrij [2008]. As discussed in Vrij [2008], the theoretical perspectives used to explain an individual's nonverbal behavior during deception also appear to be applicable in explaining the verbal content of deceptive speech. Four common theoretical perspectives are generally used in this prior research: emotions, cognitive effort, attempted control, and lack of embracement.

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

Proponents of the cognitive effort perspective argue that fabricating a lie is difficult. If a liar has little or no opportunity to prepare or rehearse, his/her verbal statements are likely to lack specific detail and instead include more general terms and little mentioning of personal experiences. Similar to the emotions perspective, this cognitive perspective implies fewer self-references and shorter statements. Thus, a liar may sound implausible and nonimmediate.

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 the statements to be a lie. Consistent with the emotions and cognitive effort theories, this perspective implies general nonspecific 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 and may use a greater number of unique words to achieve lexical diversity. In contrast, truth-tellers often repeat their information and this type of repetition leads to less lexical diversity.

Control by a speaker may also lead to a very regular or smooth speech pattern when a narrative is prepared and rehearsed in advance. In contrast, truth-tellers often adapt what they have said previously, sometimes expanding on a discussion point that they forgot to mention at an earlier point.[1] In contrast to the cognitive effort perspective, the attempted control theory

[1] Hence, to gain some basic 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 sometimes 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). To the extent that the CEO and CFO narratives are highly rehearsed, this will make it very difficult for a linguistic model to detect deception or lying about financial statements.

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 because they have not personally experienced the supposed claims. Similar to the other theories, 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 presentations and answers of the CEO and CFO are well rehearsed, the associations between specific linguistic features and deception are theoretically ambiguous. The next subsection describes specific verbal cues of deception that we include in our prediction models.

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] hypothesize that liars are less forthcoming than truth-tellers and their responses are shorter. This notion is similar to the emotions, cognitive effort, and lack of embracement perspectives, which argue that deceivers produce statements with fewer words. 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.[2]

The measurement strategy for our word categories is initially based on well-developed word lists (e.g., LIWC and WordNet). As described below, LIWC is a source for positive and negative emotions words, pronouns, certainty and tentative words, and speech hesitations. We also expand some categories by adding synonyms from a lexical database of English WordNet. To establish word categories specific to deception in the conference call setting, we examined 10 transcripts for quarters in which financial results were being subsequently restated. Based on our reading of these transcripts, we create word lists for references to general knowledge, shareholder 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. We acknowledge that our study is fundamentally exploratory in nature and some hypothesized signs in table 1 are ambiguous. The hypothesized signs are based on our assessment of the prior theoretical and

[2] We find that response length is highly positively correlated with our measure of lexical diversity, defined as the number of unique words. As a result, we include only response length in our analysis.

TABLE 1

Definitions of Linguistic-Based Variables

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 divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Bachenko, Fitzpatrick, and Schonwetter [2008], Bond and Lee [2005], DePaulo et al. [2003], Newman et al. [2003], Knapp, Hart, and Dennis [1974], Newman et al. [2003], Vrij [2008].
1st person plural pronouns	we	+	LIWC category “we”: we, us, our, etc. Simple count divided 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, Hart, and Dennis [1974], Newman, Hart, and Dennis [2003], Vrij [2008].
3rd person plural pronouns	they	+/−	LIWC category “they”: they, their, they’d, etc. Simple count divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Vrij [2008], Knapp, Hart, and Dennis [1974], Newman et al. [2003].
Impersonal pronouns	ipron	+/−	LIWC category “ipron”: it, anyone*, nobod*, etc. Simple count divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: DePaulo et al. [2003], Knapp, Hart, and Dennis [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 divided by 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 divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Vrij [2008].
Nonextreme positive emotions	posemone	−	Modified LIWC category “posemo”: love, nice, accept, etc. This LIWC category excludes extreme positive emotions words, which are listed in panel B. Simple count divided 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 divided 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 divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Adams and Jarvis [2006], Bachenko, Fitzpatrick, and Schonwetter [2008], Newman et al. [2003], Vrij [2008].

(Continued)

TABLE 1—*Continued*

Panel A: Variables, Computation, and Predicted Signs

Category	Abbreviation	Sign	Calculation
Anxiety	anx	+	LIWC category “anx”: worried, fearful, nervous, etc. Simple count divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Bachenko, Fitzpatrick, and Schonwetter [2008], Bond and Lee [2005], Knapp, Hart, and Dennis [1974], Newman et al. [2003], Vrij [2008].
Anger	anger	+	LIWC category “anger”: hate, kill, annoyed, etc. Simple count divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Bachenko, Fitzpatrick, and Schonwetter [2008], Bond and Lee [2005], Newman et al. [2003], Vrij [2008].
Swear words	swear	+	LIWC category “swear”: screw*, hell, etc. Simple count divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Bachenko, Fitzpatrick, and Schonwetter [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 divided 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].
Cognitive Process			
Certainty	certain	−	LIWC category “certain”: always, never, etc. Simple count divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Bond and Lee [2005], Knapp, Hart, and Dennis [1974], Newman et al. [2003], Vrij [2008].
Tentative	tentat	+	LIWC category “tentat”: maybe, perhaps, guess, etc. Simple count divided 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, Hart, and Dennis [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 divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Vrij [2008].
Shareholder 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 divided 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 divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample.

(Continued)

TABLE 1—*Continued*

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
Shareholder 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
Hesitations	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

(Continued)

TABLE 1—*Continued*

Panel B: Self-Constructed Word Categories	
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

This table presents definitions of the linguistic-based variables that we use to estimate classification models for deceptive instances. Panel A presents the predicted sign for each variable and provides the words included in the measure. 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 are included in these categories.

empirical psychology and linguistic literatures that have examined lying and deception.

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

We expect that deceptive executives will have fewer self-references (I) and more first-person plural pronouns (we) in their narratives. Prior studies find that 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 of general statements with deception is theoretically ambiguous, prior empirical research finds that deceivers use more generalizations. We also expect 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 to general knowledge (genknlref), which includes phrases such as "you know," "others know well," and other similar words or phrases.

Negative statements are generally recognized as indicators of a deceptive message (e.g., Adams and Jarvis [2006]). Vrij [2008] argues that lies often include statements that indicate aversion toward a person or an 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 expect that negation, anxiety, swear words, and anger are positively related to deceptive statements, whereas assent and positive emotions are negatively related to deception. We also differentiate between "extreme" and "nonextreme" words for positive and negative emotional words. In our setting, we expect that executives will use extreme positive emotional words such as "fantastic" to sound more persuasive while making a deceptive claim. To construct both categories of extreme positive and negative emotional words, we selected the words that express strong emotions from correspondingly positive emotion (posemo) and negative emotion (negemo) LIWC categories and completed the lists by adding synonyms for these words from WordNet.

The lack of embracement perspective suggests that liars lack conviction and differ from truth-tellers on the degree of certainty in their statements. 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 10 likely deceptive transcripts, we develop two categories "shareholder 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, the sign of the association between these categories and deception is theoretically ambiguous. 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 "shareholder 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, we speculate that shareholder value and value creation words may be used less when deceptive executives are concerned about future litigation associated with their actions. For example, shareholder lawsuits commonly compare the statements made by executives to their actual knowledge to show that executives were lying to shareholders. It is possible that lying executives do not use references to shareholder value or value creation during conference calls in order to avoid this legal concern.

4. Sample

Our sample is constructed using a comprehensive set of conference call transcripts provided by FactSet Research Systems Inc. We consider all available transcripts of quarterly earnings conference calls for U.S. companies over the time period from September 2003 to May 2007. A total of 29,663 transcripts were automatically parsed.

The typical conference call consists of an MD section and a Q&A section. Our initial assumption was that the MD section would be more scripted and rehearsed than the Q&A section. Intuitively, the classification power should come from cues obtained from the natural flow of speech and this should be more prevalent in the Q&A section. However, in untabulated results, we found that models based solely on the formal MD section exhibited similar ability as the Q&A section to detect serious restatements. We speculate that, if an executive also delivers scripted details using his/her own natural linguistic tendencies, we might expect to find similar results for MD sections as for Q&A sections. Based on these results, we pool the MD and Q&A sections of conference calls for developing our linguistic models.

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 speaker identifiers. In addition, transcripts have an operator (who coordinates the call) with his/her own identifier. There are no questions in MD sections and almost all corporate representatives are clearly identified in this text. There are three types of phrases that can be found in Q&A sections: the operator's introductory phrase, question posed by the caller, and answer by the executives. We assume that all answer phrases belong to corporate representatives and all question phrases belong to 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. Because the operator introduces each new outside participant, we assume that the same participant keeps asking questions until the operator introduces another participant. In addition, 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* conference call files involves the following: (1) collecting all phrases that belong to a corporate representative within an MD section; (2) finding an operator phrase that precedes the first question in a Q&A section (if this is not found, questions and answers are not recorded); (3) recording questions of the same speaker until the operator interrupts; (4) recording answers of the same speaker answering questions until a new speaker who also answers questions interrupts; and (5) requiring that a question must come after the 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 each question.

We define an instance as all phrases of a corporate representative (e.g., CEO, CFO, etc.) on a particular conference call regardless of whether they belong to MD or Q&A sections. Moreover, from the header of the *.xml* file, we can find the title of a corporate representative.[3]

We assume that CEOs and CFOs are the most likely executives to have knowledge about financial statement manipulation. Because these executives are the most common participants on conference calls, we develop separate data files for only the CEOs and CFOs. We require the length of an instance to be at least 150 words, which corresponds approximately to the mean number of words in an answer to one question. We use this constraint in order to obtain a reasonable number of words for measuring our linguistic constructs. Our CEO sample has 17,150 instances and CFO sample has 16,032 instances.

[3] 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 identification. For instance, BEDI AJAY SINGH might also be referred to as BEDI SINGH, EDWARD PARRY as ED PARRY, 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.

TABLE 2
Exchange Membership and Industry Composition

	Compustat, %	CEO, %	CFO, %
Panel A: Firms by Stock Exchange in 2005			
Nontraded Company or Security	2.42	0.27	0.29
New York Stock Exchange	27.44	45.20	46.19
NYSE Amex	6.16	1.43	1.28
OTC Bulletin Board	9.97	1.51	1.20
NASDAQ-NMS Stock Market	39.00	45.93	45.74
NYSE Arca	2.36	0.04	0.00
Other-OTC	12.64	5.62	5.30
Number of obs.	8,083	2,582	2,416
Panel B: Firms by Industry in 2005			
Mining/Construction	1.69	1.98	1.78
Food	1.53	2.01	2.11
Textiles/Print/Publish	2.90	4.65	4.47
Chemicals	1.98	2.44	2.52
Pharmaceuticals	6.34	5.96	5.22
Extractive	3.55	3.49	3.35
Durable Manufacturing	15.95	19.02	18.63
Computers	12.16	14.91	15.40
Transportation	4.69	6.16	6.46
Utilities	3.92	3.64	3.77
Retail	7.09	10.53	11.05
Financial	14.19	11.12	10.89
Insurance/RealEstate	14.49	4.69	4.68
Services	7.72	8.95	9.27
Other Industries	1.80	0.46	0.41
Number of obs.	8,281	2,582	2,416

This table presents exchange membership (panel A) and industry composition (panel B) for the sample of conference calls covering the time period from September 2003 to May 2007. Separate descriptive statistics are presented for conference calls with the CEOs and CFOs participating in the call.

The descriptive statistics for our samples are presented in table 2.[4] Approximately 90% of our firms are listed on the NYSE or NASDAQ (panel A). We find that industry distribution in our sample is close to the Compustat industry distribution (panel B). However, in untabulated results, we find that firms in our sample are significantly larger in terms of market capitalization, total assets, and sales. In addition, the firms in our sample are more profitable in terms of return on assets and profit margin and have significantly greater free cash flows than the Compustat population. These results are perhaps anticipated because larger and profitable firms would generally be expected to commit more resources to investor relation activities. The observed differences between our sample and the Compustat population limit the generalizability of our results.

[4] 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.

5. *Labeling Instances as Deceptive*

Previous research on deception and lying commonly uses controlled behavioral experiments where participants are asked to lie or to tell the truth as part of their task (e.g., Newman et al. [2003], Bond and Lee [2005], Hobson, Mayew, and Venkatachalam [2012]). This design enables the researcher to know with certainty whether a statement by the subject is deceptive or not. However, this type of experiment is fairly contrived and can differ substantially from lying in real life (i.e., there are 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 restated. We assume that these executives either intentionally manipulated the financial reports or they knew that investors were being provided with false accounting information during the conference call.

We use data from Glass, Lewis & Co. to identify quarterly reports that are restated by each firm (if the firm restates its annual financial statements, we assume that every quarter for that year is restated). These data cover restatements announced during the time period from 2003 to 2009. Glass, Lewis & Co. is a commercial company that reviews SEC filings to identify restatements that correct accounting errors (restatements related to changes in accounting principles or text corrections are excluded). Selected filters are then applied to the Glass, Lewis & Co. data to identify whether each restatement observation is "trivial" or "serious." We expect that serious accounting restatements will provide more diagnostic linguistic cues for predicting executive deception.

The most basic filter requires that a restatement involves a material weakness disclosure, a late filing, an auditor change, or a disclosure using Form 8-K. A material weakness disclosure implies that there is a deficiency in the internal controls over financial reporting. An auditor change can be a signal of deficient external monitoring. A late filing indicates 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] find that a Form 8-K filing is related to more serious restatements. We label restatements with these characteristics as NT (or "nontrivial"). The NT filter provides the weakest criterion for capturing the seriousness of a restatement.

We define the IRAI (or "irregularities or accounting issues") category as one that includes irregularities as defined in Hennes, Leone, and Miller [2008] and restatements related to revenue recognition or expense recognition issues (excluding lease accounting or stock option restatements). The choice of accounting issues is consistent with Scholz [2008], who finds that the most serious restatements for the 1997–2006 period were related to revenue recognition and core expenses errors. Revenue recognition restatements are consistently associated with more negative market returns, but this is less true for the restatements of core expenses. In particular, two core expense restatements do not exhibit negative announcement returns for a restatement. The first set of restatements is made by firms required to provide Section 404 reports and restating during 2003–2005 implementation period. The second set of restatements is made by firms that are restating lease accounting.[5] In addition, Scholz [2008] does not find a negative market reaction for stock-based and deferred compensation restatements.

We create the irregularity (IR) label following Hennes, Leone, and Miller [2008]. In particular, we extract all firm-specific news, press releases, and Form 8-K filings from LexisNexis from January 2003 to March 2011 that contain the term "restate!".[6] All news sources are searched two years before and two years after the restatement date. Exhibit 99 of Form 8-K (which contains the company's press releases) was also searched for the derivative forms of terms "irregularity" and "fraud." All other news sources, including press releases, were searched for SEC, Department of Justice formal or informal investigations, independent investigations (e.g., by audit committee or special committee), and class action lawsuits.[7] After automatic prescreening for search terms, each relevant paragraph was read to make a final judgment about the content of the news. In addition to labeling based on press releases and news, we also used the Glass, Lewis & Co. data on whether the restatement involved an SEC investigation or security class action.

Finally, we collected Accounting and Auditing Enforcement Releases (AAERs) starting from January 2003 to March 2011 from the SEC web page and searched for company name in the text of these AAERs and for variations of the terms "fraud" or "material misstatement" during the time period from one year before to three years after the restatement filing date. As with the irregularities sample, we read the leading paragraph for each AAER that satisfied the search criteria to make a final judgment regarding whether to assign the AAER label. We identify the restated quarters using the Glass, Lewis & Co. data.

The frequency of deceptive firm-quarters by year is presented in table 3. Years 2003, 2004, and 2005 have the highest rate of deceptive firm-quarters. Part of the reason for the relatively large number of restatements observed during these years is that this time period is immediately after the adoption

[5] "In a February 2005 open letter to the AICPA, then—SEC Chief Accountant Donald Nicolaisen noted the large number of companies improperly accounting for lease transactions. The letter, which Mr. Nicolaisen issued at the request of the accounting profession, laid out the SEC staff's view of the correct way to record lease transactions—a view that FASB happened to share. As it turned out, hundreds of companies simply had not been following GAAP"(Turner and Weirich [2006]).

[6] We select the following news sources from LexisNexis: SEC Form 8-K, Associated Press, Business Wire, GlobeNewswire, Marketwire, PR Newswire, and Thomson Reuters. To disregard the use of the term "restate!" in other contexts, our search statement is: "(restate! PRE/5 (per share OR statement! OR financial! OR filing! OR result!)) NOT W/255 (clawback! OR bonus! OR compensation! OR mean! OR forfeitur! OR agreement! OR plan! OR bylaw! OR incorporation! OR right!)". To avoid general news articles, we also require the company name to be mentioned in the headline.

[7] We are grateful to Andrew Leone for suggesting this approach. Hennes, Leone, and Miller [2008] report that most irregularities involve a class action lawsuit and most errors do not.

TABLE 3

Deceptive Firm-Quarters by Year for the Sample of CEOs and CFOs Narratives During Conference Calls

		NT		IRAI		IR		AAER	
	N	*N*	%	*N*	%	*N*	%	*N*	%
Panel A: CEO Sample									
2003	1,169	239	20.44	176	15.06	146	12.49	31	2.65
2004	4,210	860	20.43	567	13.47	465	11.05	87	2.07
2005	4,661	737	15.81	522	11.20	454	9.74	112	2.40
2006	5,637	427	7.57	313	5.55	260	4.61	40	0.71
2007	1,473	62	4.21	49	3.33	30	2.04	4	0.27
Total	17,150	2,325	13.56	1,627	9.49	1,355	7.90	274	1.60
Panel B: CFO Sample									
2003	1,113	226	20.31	166	14.91	140	12.58	35	3.14
2004	3,923	814	20.75	545	13.89	449	11.45	91	2.32
2005	4,389	720	16.40	512	11.67	439	10.00	105	2.39
2006	5,213	445	8.54	319	6.12	267	5.12	41	0.79
2007	1,394	56	4.02	42	3.01	28	2.01	3	0.22
Total	16,032	2,261	14.10	1,584	9.88	1,323	8.25	275	1.72

This table reports the frequency of deceptive firm-quarters by year (where the quarter was subsequently restated). The first column is the total number of firm-quarters by year. The following four columns are counts of deceptive firm-quarters under the different criteria described below, where *N* is the count of deceptive firm-quarters under a given criterion and % is the percentage of deceptive firm-quarters in the total number of firm-quarters in a given year. Nontrivial category (NT) includes restatements that involve the disclosure of a material weakness within one year before or after the restatement, or a late filing within one year before or after the restatement, or an auditor change within one year before or after the restatement, or a Form 8-K filing. Irregularities or accounting issues category (IRAI) includes restatements that involve revenue recognition or expense recognition issues (excluding lease accounting or stock option restatements) that are shown to elicit consistent negative market reaction at the time of announcement (e.g., Palmrose, Richardson, and Scholz [2004], Scholz [2008]) or irregularities according to the criteria outlined by Hennes, Leone, and Miller [2008]. Irregularities category (IR) includes restatements that involve irregularities according to one of the criteria outlined by Hennes, Leone, and Miller [2008]: (1) the firm explicitly uses derivative forms of words "fraud" or "irregularity" in the press release that discusses a restatement; (2) the firm is under formal or informal SEC or Department of Justice investigation that is related to a restatement; (3) the firm initiates independent investigation (e.g., by special committee) that is related to a restatement; or (4) there is a class action lawsuit related to a restatement. We search LexisNexis news database and company filings for the evidence on these four criteria within two years before and two years after the restatement filing date. Accounting and Auditing Enforcement Releases category (AAER) includes restatements that involve an AAER issued by the SEC within one year before and three years after the restatement filing date charging a firm or its executives with fraud or material misstatement.

of Sarbanes–Oxley and the implementation of Section 404. In addition, for the early part of our sample period, there are more years afterward for the detection of an accounting problem and subsequent restatement.[8] As should be expected, the overall percentage of deceptive firm-quarters is highest for the less restrictive NT criterion, which is 13.56% (14.10%), and lowest for the most restrictive AAER criterion, which is 1.60%(1.72%), respectively, for the samples of CEOs (CFOs). Because the time period covered by the sample has several unique institutional features such as Section 404 implementation and increased conservatism among the auditing profession, this limits our ability to generalize the results to other calendar years.

[8] 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 likely reduce the power of our statistical analysis, and thus produce conservative statistical tests.

6. *Econometric Issues*

Similar to traditional classification research, we estimate a simple binomial logistic model for our primary statistical analyses. Our outcome variable is coded as one if a conference call is labeled as deceptive and zero otherwise. There are three additional fundamental econometric choices that are necessary to generate the results. First, it is necessary to develop an appropriate cross-validation approach to estimate the *out-of-sample* classification performance of our models. Second, we need to properly evaluate whether this performance is better than chance alone. Finally, it is necessary to correctly estimate the standard errors for our panel data.[9]

6.1 EVALUATING OUT-OF-SAMPLE PERFORMANCE

To estimate the prediction error of a classifier, it is necessary to assess the out-of-sample prediction error because the in-sample prediction error is a very optimistic estimate of performance in 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 a single split may not exhibit enough variation to both fit the model and consistently estimate the out-of-sample prediction error.

To obtain a consistent estimate of the prediction error, we perform cross-validation (Efron and Tibshirani [1994], Witten and Frank [2005], Hastie, Tibshirani, and Friedman [2003)]). Specifically, the K-fold cross validation is implemented in the following manner: (1) data are split into K roughly equal samples (folds); (2) k: $k = 1, \ldots, K$ fold is fixed; (3) the model is estimated using $K - 1$ folds, ignoring the kth fold; and (4) performance of the model is evaluated using the kth fold. These steps are repeated K times where the $k = 1, \ldots, K$. 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 also implement a stratified cross-validation that forces the proportion of deceptive and nondeceptive instances in each random data split to be the same as in the original sample. When we compare the performance of different models (e.g., linguistic versus financial models), the same data split is used for each model.

6.2 EVALUATING CLASSIFICATION PERFORMANCE

There are many performance measures that are commonly employed in classification studies. The primary performance measures are accuracy (the overall rate of correctly classified instances), true positive rate (TPR; the rate of correctly classified positive instances), false positive rate (FPR;

[9] We extensively employ R software in our analyses (R Development Core Team [2005]) and specific packages: glmnet (Friedman, Hastie, and Tibshirani [2009]), ROCR (Sing et al. [2005]), xtables (Dahl [2009]), tm (Feinerer, Hornik, and Meyer [2008], Feinerer [2010]), and the modified code for clustered standard errors by Mahmood Arai.

the rate of incorrectly classified negative instances), and precision (the rate of correctly classified positive instances among all instances classified as positive). These measures are dependent on the choice of the cutoff for assigning an observation as deceptive or not deceptive.

There are two main issues with using cutoff-dependent performance measures. First, if the cutoff for the probability of a positive class is set very high, this will reduce the chance of misclassifying negative instances as positive, but at the same time will also reduce the chance of correctly classifying positive instances. Second, measures such as precision are very sensitive to the relative frequency of positive and negative instances in the sample.

To avoid the limitations related to an arbitrary choice of a cutoff and sample composition, we employ a general measure of classification performance developed using the area under the Receiver Operating Characteristics (ROC) curve that combines the TPR and the FPR in one graph. The ROC curve is the standard technique for visualizing and selecting classifiers (e.g., Fawcett [2006]). ROC graphs for two-class problems are two-dimensional graphs in which the TPR (the gain from changing a probability cutoff) is plotted on the y-axis, and the FPR (the cost from changing the probability cutoff) is plotted on the x-axis. ROC graphs do not depend on the class distribution. Independence of the class distribution implies that ROC graphs are not affected by the rare nature of positive (deceptive) instances in our study. It is possible to reduce the performance of a classification model to a single scalar by computing 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.

We test the AUC for our models against the AUC for a random classifier using the corrected resampled t-test (e.g., Nadeau and Bengio [2003], Witten and Frank [2005]). A random classifier is defined as a classification model that uses no information from the data and assigns instances to be in a positive class with a fixed probability (the AUC of a random classifier is 0.5). For our setting, the standard t-test is inappropriate because the training samples overlap in a single cross-validation run. This violates the independence of observations assumption required for the standard t-test. In addition, across the different runs of 10-fold cross validations, there will be some overlap of the testing sets. As a result, the standard t-test exhibits very high Type I error (e.g., Dietterich [1998], Nadeau and Bengio [2003]).[10] To address this statistical problem, Nadeau and Bengio [2003] suggest incorporating the confounding correlation into t-test computation for random subsampling.[11] Bouckaert and Frank [2004] use this correction for cross-validation, which is a special case of random

[10] To demonstrate the importance of this adjustment, the standard t-test is approximately 3.5 times larger than the resampled t-test for 10-fold cross-validation repeated 10 times.

[11] Nadeau and Bengio [2003 p. 249] argue that this heuristic is likely to work well when the parametric model is not overly complex and the model is robust to perturbations in the training set. The logit model satisfies these criteria.

subsampling. The corrected resampled *t*-test is

$$t = \frac{\overline{d}}{\sqrt{\left(\frac{1}{k} + \frac{n_2}{n_1}\right)\hat{\sigma}_d^2}},$$

where d is the paired difference in a performance measure (in our case, the AUC), k is the total number of cross-validation runs (in our case, $k = 100$), n_1 is the number of instances used for training, and n_2 is the number of instances used for testing (in our case, $n_2/n_1 = 1/9$). We use this corrected resampled *t*-test to test whether the AUC of our linguistic-based models is significantly different from 0.5 (the AUC of a random classifier).

6.3 ESTIMATING STANDARD ERRORS

To test the hypotheses related to estimated coefficients for specific word categories, we report two-way clustered standard errors within both executives and fiscal quarters clusters (Cameron, Gelbach, and Miller [2006], Thompson [2011], Petersen [2009], Gow, Ormazabal, and Taylor [2010]). We expect that the linguistic patterns will be correlated over time for an executive. Similarly, common macroeconomic or industry shocks that affect many firms may induce correlation across executives. For example, it is possible that quarter-specific macroeconomic shocks may cause differences in the motivation for executives to manipulate earnings. In untabulated results, we find that the standard errors estimated with clustering only on executives are very close to the standard errors estimated with clustering on both executives and fiscal quarters (i.e., clustering by fiscal quarters does not substantially change the standard errors once we cluster on executives).

7. *Results*

7.1 DESCRIPTIVE STATISTICS FOR LINGUISTIC-BASED MODELS

In order to build the classification model, we convert each instance into a vector in the space of word categories (summarized in table 1) by counting the number of words in each category. Similar to most of the prior related literature, we assume that an instance is simply a "bag-of-words" (i.e., the position of a word in a sentence is irrelevant for classification and context is ignored). We divide the word counts in each category 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 linguistic-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 3,095 (2,902) words and the mean (median) instance length for CFOs of 2,152

TABLE 4

Descriptive Statistics for the Word Categories Used for the Sample of CEO and CFO Narratives During Conference Calls

	Mean	Std Dev	25th	50th	75th	Min	Max
Panel A: CEO Sample (*N* = 17,150)							
Word Count							
wc	3,095.68	1,543.94	1,935.00	2,902.50	4,036.00	465.49	7,684.04
References							
I	37.35	17.44	24.89	34.70	47.07	7.04	92.23
we	158.43	33.86	135.24	157.76	181.10	78.17	242.88
they	15.77	9.97	8.39	13.97	21.19	0.00	47.77
ipron	168.22	36.90	142.63	167.31	193.04	84.30	259.76
genknlref	5.69	7.66	1.01	2.99	7.06	0.00	40.94
Positives/Negatives							
assent	5.91	4.52	2.66	5.05	8.09	0.00	22.18
posemone	95.81	20.16	81.61	94.03	108.07	55.08	154.90
posemoextr	21.00	9.81	13.93	19.47	26.40	4.21	53.22
negate	26.59	12.16	17.76	25.36	33.97	3.74	62.11
anx	2.61	2.71	0.59	1.93	3.79	0.00	13.23
anger	2.53	2.52	0.74	1.97	3.61	0.00	12.97
swear	0.16	0.52	0.00	0.00	0.00	0.00	3.32
negemoextr	5.47	3.86	2.69	4.78	7.57	0.00	18.38
Cognitive Mechanism							
certain	36.05	10.58	28.67	35.19	42.53	13.61	66.51
tentat	68.13	21.59	52.77	66.71	81.80	23.49	127.52
Other Cues							
hesit	0.20	0.54	0.00	0.00	0.00	0.00	3.04
shvalue	0.25	0.69	0.00	0.00	0.00	0.00	3.95
value	0.15	0.49	0.00	0.00	0.00	0.00	2.99
Panel B: CFO Sample (*N* = 16,032)							
Word Count							
wc	2,152.97	1,200.13	1,302.00	1,910.00	2,739.25	365.00	6,481.07
References							
I	16.56	9.23	9.74	15.19	21.83	1.09	45.98
we	85.54	22.92	70.53	86.06	100.68	26.54	140.65
they	4.51	4.15	1.48	3.51	6.39	0.00	20.14
ipron	89.13	26.59	70.28	88.05	106.46	32.03	159.79
genknlref	2.73	4.25	0.00	1.21	3.32	0.00	23.15
Positives/Negatives							
assent	4.15	3.55	1.61	3.36	5.82	0.00	17.36
posemone	54.02	13.48	44.48	53.24	62.53	25.09	92.06
posemoextr	8.24	5.11	4.52	7.41	11.06	0.00	24.63
negate	13.49	7.14	8.30	12.59	17.64	0.00	35.54
anx	1.42	1.92	0.00	0.78	2.10	0.00	9.39
anger	0.71	1.15	0.00	0.00	1.08	0.00	5.88
swear	0.06	0.34	0.00	0.00	0.00	0.00	2.74
negemoextr	2.01	2.04	0.00	1.55	3.05	0.00	9.08
Cognitive Mechanism							
certain	18.48	6.96	13.63	17.93	22.80	3.92	38.27
tentat	41.00	13.41	31.37	39.87	49.21	13.84	79.39
Other Cues							
hesit	0.10	0.37	0.00	0.00	0.00	0.00	2.26
shvalue	0.06	0.26	0.00	0.00	0.00	0.00	1.62
value	0.03	0.15	0.00	0.00	0.00	0.00	1.08

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

(1,910) words. When both executives speak at a conference call, CEO instances contain a greater number of words than CFO instances for approximately 70% of firms. For the reference category, impersonal pronouns have the highest word count and references to the general knowledge have the lowest word count. The largest word count for the positives/negatives category is nonextreme positive emotion words with negations being the second largest category. As might be expected given the public nature of conference calls, the category of swear words has the lowest word count. Both executives use almost twice as many tentative words as words expressing certainty. Finally, there are very few hesitations and low usage of shareholder value and value creation phrases.

7.2 OUT-OF-SAMPLE CLASSIFICATION PERFORMANCE OF LINGUISTIC-BASED MODELS

The classification models based on linguistic cues of CEOs and CFOs perform significantly better than a random classifier by about 6%–16%, depending on the deception labeling criteria (table 5). The model has the lowest classification power for the NT label with the gain over a random classifier of about 6% for CEOs and 8% for CFOs. As discussed in section 5, this result may be caused by measurement error in the NT proxy for serious deceptions. The model has the greatest classification power for the most restrictive labeling criterion (AAER), with the gain over a random classifier of about 13% for CEOs and 16% for CFOs. Thus, classification power of the linguistic model increases as the labeling criteria for deception become more restrictive.

As discussed in section 6, classification performance measures other than AUC (e.g., TPR, FPR, precision, and accuracy) are cutoff-dependent. The outcome of our logit model is the probability that an instance is deceptive, and the classification rule is that observations with probabilities higher than the chosen cutoff are assigned to the deceptive class. To provide an assessment for the cutoff-dependent measures, we present results for cutoffs at 50th, 70th, and 90th percentile of the predicted probability of a deceptive class in the pooled sample. As the cutoff level is increased, both the TPR (the benefit of the classification) and the FPR (the cost of the classification) fall. However, by increasing a cutoff, there is also a gain in precision (i.e., there is a greater proportion of true deceptive instances among those instances that are classified by the algorithm as deceptive).[12]

7.3 LINGUISTIC CUES AND THE LIKELIHOOD OF DECEPTION

The estimated associations between linguistic cues and the likelihood of deception are reported in table 6 for CEOs (panel A) and CFOs

[12] Precision cannot be compared across restatement labels because this performance metric depends on the relative proportion of positives and negatives in the sample. In particular, we have the lowest rate of deceptive instances under the AAER criterion and the lowest precision.

TABLE 5

Classification Performance of Linguistic-Based Prediction Models for CEO and CFO Narratives During Conference Calls

	NT	IRAI	IR	AAER
Panel A: CEO Sample				
Sample Composition				
Total firm-quarters	17,150	17,150	17,150	17,150
Deceptive firm-quarters	2,325	1,627	1,355	274
Deceptive firm-quarters(%)	13.56	9.49	7.90	1.60
Performance				
AUC(corrected *t*-test vs. 50%)	56.64(10.77)	58.03(11.29)	59.72(10.79)	63.03(6.99)
Cutoff at the 50th Percentile				
TPR in %	58.09	59.58	62.21	66.79
FPR in %	48.74	48.88	48.93	49.70
Precision in %	15.75	11.33	9.83	2.13
Accuracy in %	52.19	51.92	51.95	50.56
Cutoff at the 70th Percentile				
TPR in %	38.29	40.34	42.89	44.32
FPR in %	28.82	29.01	29.02	29.77
Precision in %	17.26	12.74	11.26	2.36
Accuracy in %	66.72	68.08	68.76	69.82
Cutoff at the 90th Percentile				
TPR in %	14.35	16.47	19.07	24.69
FPR in %	9.45	9.34	9.20	9.92
Precision in %	19.33	15.69	15.14	3.93
Accuracy in %	80.22	83.62	85.13	89.03
Panel B: CFO Sample				
Sample Composition				
Total firm-quarters	16,032	16,032	16,032	16,032
Deceptive firm-quarters	2,261	1,584	1,323	275
Deceptive firm-quarters(%)	14.10	9.88	8.25	1.72
Performance				
AUC(corrected *t*-test vs. 50%)	58.06(12.43)	58.43(11.94)	59.64(12.38)	66.75(10.20)
Cutoff at the 50th Percentile				
TPR in %	59.41	60.76	62.24	71.70
FPR in %	48.27	48.91	48.78	49.47
Precision in %	16.81	11.99	10.29	2.47
Accuracy in %	52.81	52.05	52.13	50.89
Cutoff at the 70th Percentile				
TPR in %	38.76	39.94	41.81	54.40
FPR in %	28.61	28.97	28.87	29.60
Precision in %	18.21	13.13	11.53	3.11
Accuracy in %	66.79	67.96	68.71	70.12
Cutoff at the 90th Percentile				
TPR in %	15.61	15.44	16.40	27.30
FPR in %	9.22	9.49	9.52	9.71
Precision in %	21.80	15.22	13.47	4.71
Accuracy in %	80.18	83.09	84.37	89.21

This table reports classification performance of the logit models that use the 19 linguistic-based variables (defined in table 1) for CEOs (panel A) and CFOs (panel B) to predict deceptive instances under NT, IRAI, IR, and AAER criteria (defined in table 3). We compute means over 100 cross-validation runs of out-of-sample performance measures: AUC (the area under ROC) in percentages, TPR (the percentage of correctly classified deceptive instances), FPR (the percentage of incorrectly classified truthful instances), precision (the percentage of actual deceptive instances among those classified by the algorithm as deceptive), and accuracy (the percentage of correctly classified instances). The classification rule is to assign all instances above a given cutoff for the probability of deception to the deceptive class and below to the truthful class. The cutoffs are at the 50th, 70th, and 90th percentiles of the predicted probability of deception estimated using pooled data. “Corrected *t*-test vs. 50%” is the value of the corrected resampled *t*-statistic testing the null hypothesis of the mean AUC being equal to 50%, which is the AUC of a random classifier (e.g., Nadeau and Bengio [2003], Bouckaert and Frank [2004]). Explanatory variables are winsorized at 1 and 99 percentiles.

TABLE 6

Logit Linguistic-Based Prediction Models for CEO and CFO Narratives During Conference Calls

Panel A: CEO Sample

		NT	IRAI	IR	AAER
		Word Count			
wc‡	±	1.01	1.04	1.16	1.04
		(0.10)	(0.13)	(0.15)	(0.24)
		References			
I	−	0.95	0.89	0.87	0.87
		(0.07)	(0.09)	(0.10)	(0.18)
we	+	0.99	0.92	0.95	1.07
		(0.04)	(0.05)	(0.05)	(0.12)
they	±	1.06	1.10	0.98	0.50**
		(0.12)	(0.16)	(0.16)	(0.15)
ipron	±	0.94	0.97	0.96	1.14
		(0.04)	(0.05)	(0.06)	(0.12)
genknlref	±	1.91***	1.96***	1.99***	1.98**
		(0.33)	(0.33)	(0.36)	(0.64)
		Positives/Negatives			
assent	−	1.10	1.16	1.20	0.36
		(0.28)	(0.36)	(0.43)	(0.28)
posemone	−	0.88**	0.94	0.93	0.97
		(0.05)	(0.07)	(0.08)	(0.16)
posemoextr	±	1.20	1.62***	1.99***	3.51***
		(0.16)	(0.25)	(0.33)	(1.26)
negate	+	0.92	0.86	0.87	1.24
		(0.11)	(0.13)	(0.15)	(0.43)
anx	+	0.38**	0.34**	0.25***	0.08**
		(0.16)	(0.14)	(0.11)	(0.08)
anger	+	0.97	1.16	1.32	0.57
		(0.35)	(0.55)	(0.70)	(0.66)
swear†	+	0.97	0.95	0.94	1.03
		(0.07)	(0.06)	(0.07)	(0.15)
negemoextr	+	0.99	0.84	0.88	0.83
		(0.26)	(0.31)	(0.33)	(0.66)
		Cognitive Mechanism			
certain	−	1.16	0.90	0.88	0.75
		(0.13)	(0.13)	(0.14)	(0.18)
tentat	+	0.96	0.96	1.00	0.99
		(0.07)	(0.08)	(0.10)	(0.19)
		Other Cues			
hesit†	±	1.05	1.04	1.11*	0.99
		(0.05)	(0.05)	(0.06)	(0.16)
shvalue†	±	0.91**	0.90*	0.88**	0.95
		(0.04)	(0.05)	(0.06)	(0.12)
value†	±	0.90	0.87	0.83*	1.11
		(0.07)	(0.08)	(0.09)	(0.17)
Total firm-quarters		17,150	17,150	17,150	17,150
Deceptive firm-quarters		2,325	1,627	1,355	274
Area under the ROC curve		0.58	0.59	0.61	0.66
Log-likelihood value		−6,732.51	−5,294.87	−4,638.95	−1,353.13
Pseudo R-squared		0.011	0.016	0.021	0.037

(Continued)

TABLE 6—*Continued*

Panel B: CFO Sample

		NT	IRAI	IR	AAER
		Word Count			
wc‡	±	1.16^{*}	1.29^{***}	1.29^{***}	1.74^{***}
		(0.09)	(0.11)	(0.12)	(0.29)
		References			
I	−	0.80^{**}	0.90	0.94	0.65
		(0.07)	(0.10)	(0.12)	(0.17)
we	+	0.99	1.01	1.04	1.06
		(0.04)	(0.04)	(0.05)	(0.12)
they	±	0.67^{***}	0.67^{**}	0.61^{***}	0.30^{***}
		(0.10)	(0.12)	(0.11)	(0.10)
ipron	±	0.93^{**}	0.90^{**}	0.88^{***}	1.21^{**}
		(0.03)	(0.04)	(0.04)	(0.10)
genknlref	±	2.05^{***}	1.95^{***}	2.04^{***}	2.47^{***}
		(0.41)	(0.40)	(0.45)	(0.77)
		Positives/Negatives			
assent	−	1.09	1.00	1.02	0.49
		(0.20)	(0.26)	(0.27)	(0.29)
posemone	−	0.83^{***}	0.89	0.85^{*}	1.05
		(0.05)	(0.07)	(0.07)	(0.16)
posemoextr	±	0.97	1.02	1.27	1.84
		(0.16)	(0.19)	(0.26)	(0.73)
negate	+	1.23^{*}	1.30^{*}	1.40^{**}	1.46
		(0.16)	(0.20)	(0.24)	(0.62)
anx	+	0.70	0.68	0.73	0.25
		(0.25)	(0.32)	(0.40)	(0.32)
anger	+	0.52	0.43	0.44	0.63
		(0.28)	(0.32)	(0.34)	(0.82)
swear†	+	1.07	1.11	1.01	1.52^{*}
		(0.14)	(0.16)	(0.16)	(0.38)
negemoextr	+	1.37	1.82	1.85	6.96^{**}
		(0.45)	(0.73)	(0.80)	(5.94)
		Cognitive Mechanism			
certain	−	1.45^{***}	1.31^{**}	1.30^{**}	1.07
		(0.14)	(0.16)	(0.17)	(0.30)
tentat	+	1.17^{**}	1.20^{**}	1.22^{**}	1.08
		(0.08)	(0.10)	(0.12)	(0.22)
		Other Cues			
hesit†	±	1.03	0.96	0.93	0.95
		(0.08)	(0.09)	(0.09)	(0.19)
shvalue†	±	0.76^{*}	0.91	0.89	0.66
		(0.11)	(0.16)	(0.19)	(0.24)
value†	±	1.04	0.70	0.56	0.63
		(0.24)	(0.22)	(0.20)	(0.33)

(Continued)

(panel B). Most of the coefficients reported in table 6 are the factors by which the odds of an instance being deceptive change when the number of words in a category increases by 1% of the median instance length. However, for word count, the reported coefficient is the change in odds for an

TABLE 6—*Continued*

Panel B: CFO Sample

	NT	IRAI	IR	AAER
Total firm-quarters	16,032	16,032	16,032	16,032
Deceptive firm-quarters	2,261	1,584	1,323	275
Area under the ROC curve	0.59	0.6	0.61	0.69
Log-likelihood value	−6,426.29	−5,090.9	−4,482.84	−1,315.32
Pseudo *R*-squared	0.015	0.015	0.018	0.054

This table presents the estimation results for logit models that use the 19 linguistic-based variables (defined in table 1) for CEOs (panel A) and CFOs (panel B) to predict deceptive instances under NT, IRAI, IR, and AAER criteria (defined in table 3). The reported coefficients are the factors by which the odds of a deceptive instance are increased if the number of words in a particular category increases by 1% of the median instance length. Specifically, for CEOs, we report $e^{29\hat{\beta}}$ (i.e., 1% of 2,902) and for CFOs, we report $e^{19\hat{\beta}}$ (i.e., 1% of 1,910). For the word count *wc*‡, we present the effect of increasing the instance length by the median instance length (i.e., $e^{2,902\hat{\beta}}$ for CEOs and $e^{1,910\hat{\beta}}$ for CFOs). For *swear*†, *hesit*†, *shvalue*†, and *value*†, we report the effect of increasing the number of words in the corresponding category by one word (i.e., $e^{\hat{\beta}}$). The standard errors for $e^{n\hat{\beta}}$ are presented in parentheses. We compute the two-way clustered standard errors by executives and fiscal quarters following Cameron, Gelbach, and Miller [2006] and convert them to the standard errors for $e^{n\hat{\beta}}$ by the delta method (i.e., let $se(\hat{\beta})$ be a standard error estimate for $\hat{\beta}$, then $ne^{n\hat{\beta}}se(\hat{\beta})$ is the standard error estimate for $e^{n\hat{\beta}}$, where n is a constant). Estimates of intercepts are omitted. Explanatory variables are winsorized at 1 and 99 percentiles. *, **, and *** denote whether the factors that correspond to the estimated coefficients are significant at 10%, 5%, and 1% level (two tailed test).

increase in the length of an instance by median length. For rare categories (e.g., swear words, hesitations, references to shareholders value, and value creation), the reported coefficient is the change in odds for a one-word change in these categories. If the reported coefficient is greater than (less than) 1, this implies that the associated estimated logit coefficient is greater than (less than) zero.[13]

7.3.1. Similarities in Linguistic Cues Between CEOs and CFOs. Several linguistic cues are significantly associated with the likelihood of deception for both executives (table 6). For example, reference to the general knowledge category (e.g, "you know," "everybody knows") is a very strong predictor that an instance is deceptive for both CEOs and CFOs. For CEOs, the effect of an increase in general knowledge phrases by 29 increases the odds of deception by a factor of 1.91 for the NT to 1.98 for the AAER criterion.

[13] The models in table 6 can be easily used to estimate the likelihood of deception in future research. Specifically, the logit model for predicting the probability of a deceptive narrative $P(Y_i = 1|X_i)$ using word categories X_i:

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

where β are raw coefficient estimates from the logit model. Note that, for the majority of word categories, table 6 reports the multipliers for the odds of deception as the number of words in a particular category increases by 1% of the median instance length. Specifically, for the word category j, the estimates in table 6 are of the form $e^{n\beta_j}$, where $n = 29$ for CEOs and $n = 19$ for CFOs. In contrast, for rare verbal cues such as swear words, hesitations, shareholder value, and value creation, $n = 1$ for both CEOs and CFOs. Let $\gamma_j = e^{n\beta_j}$ be the number reported in table 6, the corresponding raw coefficient $\beta_j = ln(\gamma_j)/n$.

For CFOs, the effect of an increase by 19 in general knowledge phrases is between 2.05 for the NT to 2.47 for the AAER criterion.[14] Furthermore, some findings are more consistent in terms of the statistical significance across criteria for one executive but not for the other. For example, both executives use significantly fewer nonextreme positive emotion words and fewer third person plural pronouns in deceptive narratives with stronger evidence found for CFOs. Finally, there is a modest negative association between shareholder value phrases and the likelihood of deception for CEOs and CFOs, although these phrases are more indicative of nondeceptive narratives for CEOs.

7.3.2. Differences in Linguistic Cues Between CEOs and CFOs. There are also a variety of substantive differences in the significant linguistic cues related to deception by CFOs and CEOs.[15] There is strong evidence for a positive association between word count and the probability of deception for CFOs, but not for CEOs. However, the increase in the length of an instance has to be substantial (1,910 words) to increase the odds of deception by a factor that ranges from 1.16 for the NT to 1.74 for the AAER criterion. Among reference categories, first person singular pronouns and impersonal pronouns have a significant negative association with the probability of deception for CFOs under less restrictive criteria.

There is also a positive association between deception and negation words for CFOs, but not for CEOs. If the number of negation words increases by 19, the odds of deception increase by a factor that ranges from 1.23 for the NT to 1.40 for the IR criterion. Furthermore, for the most restrictive deception criterion (AAER), swear and extreme negative emotion words are predictive of deception for CFOs. However, a similar result is not found for CEOs. In contrast to an expected positive association, there is a strong negative association with anxiety words for CEOs, but not for CFOs. In addition, only for CEOs, there is a positive association of extreme positive emotion words with the likelihood of deception. As the number of words in the extreme positive emotion category increases by 29, the

[14] The reference to general knowledge category (e.g., "you know", "everybody knows") can measure the general nature of the statement, a filler-type phrase, or both depending on its linguistic role in the sentence. The positive association with deception of these phrases may be related to the absence of a very detailed script of the conference call because both filler-type phrases and generalizations are likely to be used when a speaker does not have enough substance to convey to the listener. Alternatively, these phrases may serve a purely functional purpose by establishing a conversational (personal) link between the speaker and the audience.

[15] Consistent with this observation, we also find that the CEO and CFO models make somewhat different predictions regarding whether a conference call is deceptive. Using a cutoff probability at the 90th percentile of the predicted probability of deception, we find that the two models agree between 66% and 78% of the time for deceptive classifications and approximately 85% of the time for nondeceptive classifications, depending on the label used for deception.

odds of deception increase by a factor that ranges from 1.62 for the IRAI to 3.51 for the AAER criterion. Finally, deceptive CFOs use more tentative and certainty words, whereas CEOs tend to use more hesitations and fewer value creation phrases in deceptive narratives according to the IR criterion.

7.3.3. Linguistic Cues and Theories of Deception. Given the exploratory nature of our analysis, we do not attempt to draw definitive conclusions about the descriptive validity of alternative deception theories. However, it is interesting to compare the results in table 6 with the theoretical perspectives described in section 3 to speculate about the deception theories that are consistent with the linguistic cues used by CEOs and CFOs. Although we are not aware of any rigorous studies on the personality or linguistic differences between CEOs and CFOs, it is possible that the personal characteristics of people attaining the CEO position are quite different from those for a CFO. Thus, different theories of deception are likely to be descriptive for CEOs and CFOs.

The significant linguistic cues for the CEO are not consistent with the emotions perspective, which theorizes that deceivers feel guilt and are afraid of being caught. In this theoretical setting, executive deceptive speech should convey more negative emotions. However, we do not find that negative emotion words have a positive association with deception for CEOs. In contrast, we find that anxiety words have a consistent negative association and extreme positive emotion words a very strong positive association when CEOs are being deceptive. This finding can be attributed to CEOs deliberately controlling the content of their speech, which is consistent with the attempted control perspective.

The significant linguistic cues for the CFO seem to be most consistent with the emotions theory of deception. Specifically, CFOs are more negative as evidenced by their greater number of negations and, under the AAER criterion, swear and extreme negative emotion words. In addition, they have fewer nonextreme positive emotion words in deceptive narratives.

Finally, there is some evidence for the attempted control perspective for both CEOs and CFOs. The attempted control theory argues that a deceiver deliberately controls the content of the narrative or extensively prepares beforehand. If the extensive use of shareholder value phrases in the call that discusses results of misstated financial statements imposes greater litigation risk on the executives, the executives might deliberately restrain themselves from the use of these phrases. Consistent with this idea, we find evidence that CEOs (and to a lesser extent CFOs) use fewer shareholder value phrases in the deceptive narratives. For CFOs, we find greater use of certainty words, longer passages, and fewer generalizations as reflected in the number of impersonal pronouns. These results are also consistent with the control theory of deception.

7.4 MODELS WITH LINGUISTIC CUES VERSUS MODELS WITH FINANCIAL VARIABLES

Although our linguistic models perform better than a random classification, a more interesting benchmark comparison is the performance of the traditional financial-variables-based models used to predict earnings management (e.g., Jones, Krishnan, and Melendrez [2008], Dechow et al. [2011], Price, Sharp, and Wood [2011], and Daines, Gow, and Larcker [2010]). We consider five different benchmark models: (1) modified Jones model discretionary accruals as in Dechow, Sloan, and Sweeney [1995]; (2) performance-matched discretionary accruals as in Kothari, Leone, and Wasley [2005]; (3) the accounting score developed by the commercial firm Audit Integrity Inc.; (4) Model 1 from Dechow et al. [2011][16]; and (5) the model developed in Beneish [1999].[17]

The variables used in the traditional accounting prediction models are defined in table 7. For all five models, we include all firms regardless of their industry membership.[18] We use Compustat Point in Time Historical data rather than Compustat Quarterly because we want the originally reported (not ex post adjusted) financial variables for our model estimation. As shown in table 8, the mean (median) of quarterly discretionary accruals estimated using the modified Jones model (denoted as *mnda*) is approximately 2.4% (2.0%) of lagged total assets. The mean (median) of performance-matched discretionary accruals (denoted as *pmnda*) is −0.7% (−0.6%) of lagged total assets.[19] The last two models, model 1 from Dechow et al. [2011] and the Beneish [1999] model, include two measures of total accruals. Model 1 from Dechow et al. [2011] uses total accruals as in Richardson et al. [2005], which sums the change in noncash working capital, the change in net noncurrent operating assets, and the change in net financial assets. The mean (median) value of these accruals (denoted as $rsst\_acc$), scaled by average total assets, is 1.3% (0.7%). The Beneish

[16] Dechow et al. [2011] propose three models for predicting severe accounting misstatements (AAERs). Model 1 includes only financial statement variables. Model 2 adds off-balance sheet and nonfinancial variables to model 1, and model 3 adds stock-market-based variables to model 2. Although models 2 and 3 include richer sets of explanatory variables, the classification power of these models according to Dechow et al. [2011] (table 7, Panels B and C) does not differ much from the performance of model 1. Thus, we only use model 1 in our analysis.

[17] 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 completely applicable to quarterly data.

[18] However, discretionary accruals models might not be applicable for financial and utility firms.

[19] The performance-matched discretionary accruals are computed by subtracting the mean of discretionary accruals estimated using the Jones model for firms in the same two-digit SIC industry code and the same fiscal quarter with return on assets being within a 1% interval of the firm's return on assets.

TABLE 7

Definitions of the Variables Used in Financial Prediction Models

Category	Abbreviation	Calculation
		Proxies for Accruals Manipulation
Modified Jones model discretionary accruals	mnda	The residuals from the cross-sectional regressions for every two-digit SIC code and fiscal quarter of total accruals measure on constant term, reciprocal of ATQ_{t-1}, $\Delta SALEQ_t - \Delta RECTQ_t$ (Jones model discretionary accruals use $\Delta SALEQ_t$), and $PPENTQ_t$. All variables are scaled by ATQ_{t-1}. The total accruals are measured following Hribar and Collins [2002] as $IBCQ_t - (CFOQ_t - XIDOCQ_t)$, if missing as $NIQ_t - OANCFQ_t$ or as implied by the balance sheet approach.
Performance-matched discretionary accruals	pmnda	The difference between Jones model discretionary accruals for firm i and the mean Jones model discretionary accruals for the matched firms where the matching is performed based on two-digit SIC code, fiscal quarter, and ROA_t for a matched firm being within 1% interval of firm i's ROA_{it}. Here, the ROA_t is computed following Kothari, Leone, and Wasley [2005] as NIQ_t/ATQ_{t-1}.
Accounting score	ai_acct_score	The commercial accounting score by Audit Integrity Inc. The lowest score (equal to 1) corresponds to the high accounting risk and the highest score (equal to 5) corresponds to the low accounting risk.
Accruals as in Richardson et al. [2005]	rsst_acc	The accruals computed following Richardson et al. [2005] as the sum of the change in noncash. working capital, the change in net noncurrent operating assets, and the change in net financial assets scaled by seasonal average of total assets (i.e., $savgATQ = (ATQ_t + ATQ_{t-4})/2$). The formula simplifies to $(((ATQ_t - LTQ_t - PSTKQ_t) - (CHEQ_t - IVST_t)) - ((ATQ_{t-1} - LTQ_{t-1} - PSTKQ_{t-1}) - (CHEQ_{t-1} - IVST_{t-1})))/savgATQ_t$, where $IVST$ is available only at the annual frequency (we assume that it does not change within a year).
Total accruals to total assets	tata	The measure of total accruals computed following Beneish [1999] as $((ACTQ_t - ACTQ_{t-1}) - (CHEQ_t - CHEQ_{t-1}) - (LCTQ_t - LCTQ_{t-1}) - (DD1_t - DD1_{t-1}) - (TXPQ_t - TXPQ_{t-1}) - DPQ_t)/ATQ_t$, where $DD1$ is available only on the annual frequency (we assume that it does not change within a year).
		Controls from Correia [2009]
Actual issuance	capmkt	An indicator variable coded 1 if the firm issues securities or long-term debt ($SSTKQ > 0$ or $DLTISQ > 0$) and 0 otherwise.
Market capitalization	nmcap	$(CSHOQ_t \cdot PRCCQ_t)/ATQ_{t-1}$.
Free cash flow	nfcf	$(CFOQ_t - CAPXQ\_Mean_t)/ATQ_{t-1}$, where we compute $CAPXQ\_Mean_t$ over 12 quarters requiring at least three non-missing observations.
Seasonal change in cash sales	sch_cs	$((SALEQ_t - \Delta RECTQ_t) - (SALEQ_{t-4} - \Delta RECTQ_{t-4}))/(SALEQ_{t-4} - \Delta RECTQ_{t-4})$.

(Continued)

TABLE 7—*Continued*

Category	Abbreviation	Calculation
		Model 1 from Dechow et al. [2011]
Accruals as in Richardson et al. [2005]	rsst_acc	Defined above.
Seasonal change in receivables	sch_rec	$(RECTQ_t - RECTQ_{t-4})/savgATQ$, where $savgATQ = (ATQ_t + ATQ_{t-4})/2$.
Seasonal change in inventory	sch_inv	$(INVTQ_t - INVTQh_{t-4})/savgATQ$, where $savgATQ = (ATQ_t + ATQ_{t-4})/2$.
Soft assets	soft_assets	$(ATQ_t - PPENTQ_t - CHEQ_t)/ATQ_{t.}$
Seasonal change in cash sales	sch_cs	Defined above.
Seasonal change in ROA	sch_roa	$ROA_t - ROA_{t-4}$.
Actual issuance	capmkt	Defined above.
		Beneish [1999] Model
Seasonal days' sales in receivables index	sdsri	$(RECTQ_t/SALEQ_t)/(RECTQ_{t-4}/SALEQ_{t-4})$.
Seasonal gross margin index	sgmi	$((SALEQ_{t-4} - COGSQ_{t-4})/SALEQ_{t-4})/((SALEQ_t - COGSQ_t)/SALEQ_t)$.
Seasonal asset quality index	saqi	$(1 - (ACTQ_t + PPENTQ_t)/ATQ_t)/(1 - (ACTQ_{t-4} + PPENTQ_{t-4})/ATQ_{t-4})$.
Seasonal sales growth index	ssgi	$SALEQ_t/SALEQ_{t-4}$.
Seasonal depreciation index	sdepi	$((DPQ_{t-4} - AMQ_{t-4})/(DPQ_{t-4} - AMQ_{t-4} + PPENTQ_{t-4}))/((DPQ_t - AMQ_t)/(DPQ_t - AMQ_t + PPENTQ_t))$, where $AMQ = AM/4$ is derived from the annual item AM.
Seasonal sales, general, and administrative expenses	ssgai	$(XSGAQ_t/SALEQ_t)/(XSGAQ_{t-4}/SALEQ_{t-4})$.
Seasonal leverage index	slvgi	$((DLTTQ_t + LCTQ_t)/ATQ_t)/((DLTTQ_{t-4} + LCTQ_{t-4})/ATQ_{t-4})$.
Total accruals to total assets	tata	Defined above.

Compustat XPF data items: ATQ is Assets—Total, SALEQ is Sales/Turnover (Net), RECTQ is Receivables—Total, PPENTQ is Property Plant and Equipment—Total (Net), IBCQ is Income Before Extraordinary Items, XIDOCQ is Extraordinary Items and Discontinued Operations (Statement of Cash Flows), NIQ is Net Income (Loss), OANCFQ is Operating Activities—Net Cash Flow, LTQ is Liabilities—Total, PSTKQ is Preferred/Preference Stock (Capital)—Total, CHEQ is Cash and Short-Term Investments, IVST is Short-Term Investments—Total, ACTQ is Current Assets—Total, LCTQ is Current Liabilities—Total, DD1 is Long-Term Debt Due in One Year, TXPQ is Income Taxes Payable, DPQ is Depreciation and Amortization, SSTKQ is Sale of Common and Preferred Stock, DLTISQ is Long-Term Debt—Issuance, CSHOQ is Common Shares Outstanding, PRCCQ is Price-Close-Quarter, COGSQ is Cost of Goods Sold, AM is Amortization of intangibles, XSGAQ is Selling General and Administrative Expense, CAPXQ is Capital Expenditures. All variables are computed using originally reported (unrestated) quarterly data from Compustat Point in Time Historical data. The final variables are winsorized at 1 and 99 percentile.

TABLE 8

Descriptive Statistics for the Variables Used in Financial Prediction Models

	Mean	Std Dev	25th	50th	75th	Min	Max
Panel A: Modified Jones Model Discretionary Accruals (N = 9,593)							
mnda	0.024	0.082	−0.012	0.020	0.058	−1.370	0.629
capmkt	0.962	0.192	1.000	1.000	1.000	0.000	1.000
nmcap	1.703	1.610	0.737	1.239	2.110	0.054	16.146
nfcf	0.023	0.085	−0.010	0.024	0.065	−0.550	0.257
sch_cs	0.171	0.582	0.006	0.102	0.233	−2.480	5.442
Panel B: Performance-Matched Discretionary Accruals (N = 9,220)							
pmnda	−0.007	0.075	−0.036	−0.006	0.021	−1.358	0.551
capmkt	0.963	0.190	1.000	1.000	1.000	0.000	1.000
nmcap	1.675	1.544	0.744	1.232	2.081	0.054	16.146
nfcf	0.024	0.082	−0.009	0.025	0.064	−0.550	0.257
sch_cs	0.169	0.573	0.007	0.102	0.231	−2.480	5.442
Panel C: Audit Integrity Accounting Score (N = 8,647)							
ai_acct_score	2.899	1.299	2.000	3.000	4.000	1.000	5.000
capmkt	0.964	0.186	1.000	1.000	1.000	0.000	1.000
nmcap	1.617	1.565	0.694	1.162	1.993	0.054	16.146
nfcf	0.024	0.083	−0.007	0.025	0.064	−0.550	0.257
sch_cs	0.174	0.617	0.004	0.100	0.235	−2.480	5.442
Panel D: Dechow, Ge, Larson, and Sloan [2010] Model 1 (N = 10,079)							
rsst_acc	0.013	0.076	−0.011	0.007	0.029	−0.311	0.502
sch_rec	0.016	0.043	−0.001	0.009	0.027	−0.157	0.250
sch_inv	0.011	0.034	0.000	0.001	0.016	−0.120	0.197
soft_assets	0.589	0.234	0.417	0.615	0.776	0.042	0.982
sch_cs	0.163	0.643	−0.001	0.098	0.230	−2.480	5.442
sch_roa	0.001	0.035	−0.005	0.001	0.007	−0.194	0.257
capmkt	0.908	0.289	1.000	1.000	1.000	0.000	1.000
Panel E: Beneish(1999) Model (N = 7,561)							
sdsri	1.044	0.319	0.929	1.004	1.089	0.199	4.018
sgmi	0.992	0.435	0.954	0.999	1.045	−3.243	4.729
saqi	1.242	1.396	0.900	0.995	1.116	0.099	16.542
ssgi	1.161	0.330	1.027	1.105	1.220	0.274	4.646
sdepi	1.063	0.457	0.914	1.002	1.110	−0.435	4.304
ssgai	1.014	0.241	0.929	0.997	1.065	0.327	2.686
slvgi	1.031	0.325	0.898	0.978	1.081	0.254	3.461
tata	−0.009	0.051	−0.026	−0.009	0.010	−0.243	0.203

This table reports descriptive statistics for the variables used in the binomial logit models that only include financial variables (defined in table 7). All variables are winsorized at 1 and 99 percentiles.

[1999] model's proxy for accrual manipulation is total accruals defined as the change in noncash working capital less depreciation (denoted as *tata*). The mean (median) of this accruals measure is −0.9% (−0.9%) of current total assets.

We also consider the commercial accounting score produced by Audit Integrity Inc.[20] Three prior studies have considered the ability of the broader

[20] The accounting score attempts to measure the likelihood of misrepresentation in the company financial statements. This score includes measures that can be subdivided into three categories: revenue recognition, expense recognition, and asset and liability valuation.

Audit Integrity's governance index or just of the accounting score to predict restatements, that is, Daines, Gow, and Larcker [2010], Correia [2009], and Price, Sharp, and Wood [2011]. The mean (median) value for the accounting score is 2.8 (3.0) where the score of 1 represents a high risk of accounting misstatement and the score of 5 represents a low risk of accounting misstatement. Similar to Correia [2009], we incorporate selected controls in the prediction models using estimates of discretionary accruals or the accounting score. Our control variables are an indicator variable for whether the firm issued securities, market capitalization scaled by lagged total assets, free cash flows scaled by lagged total assets, and seasonal growth in cash sales.[21]

We perform pairwise tests of the AUC measures for the models based on CEOs' (CFOs') narratives and five alternative accounting models. Similar to our earlier methodological approach, the pairwise comparisons use 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 two alternative models using the same data for estimation and the same data for out-of-sample testing. We also use the conservative corrected resampled *t*-tests to assess the significance of the differences between the AUC measures.

In table 9, we find that models based on linguistic cues from CEO narratives are never statistically dominated by models using only financial variables. However, for the less restrictive criteria (NT, IRAI, and IR), some financial-variables-based models perform statistically worse than the linguistic-based models. For example, the model that uses modified Jones model discretionary accruals performs significantly worse than the linguistic-based model for the IRAI (57.46% versus 53.68%) and the IR (59.24% versus 54.58%) criteria. The model based on performance-matched discretionary accruals performs significantly worse for the IR (59.05% versus 55.94%) criterion. Similar results hold for model 1 from Dechow et al. [2011] and the Beneish [1999] model, which perform significantly worse than the linguistic-based model under the NT, IRAI, and IR criteria.

Similar to CEO model comparisons, some accounting models are dominated by the linguistic-based models for CFOs. Specifically, the model that uses performance-matched discretionary accruals performs significantly worse under the NT (57.53% versus 54.88%) and the IRAI (58.91% versus 54.62%) criteria. Similarly, the modified Jones model discretionary accruals, model 1 from Dechow et al. [2011] and the Beneish [1999] model perform statistically worse than the linguistic-based model for CFOs under the NT, IRAI, and IR criteria.

[21] The inclusion of these control variables will increase the AUC for these three models beyond the AUC attributable to the accounting measure alone. Thus, the comparative tests between linguistic and accounting-based models will be conservative.

TABLE 9

Classification Performance of Linguistic-Based and Financial Variables–Based Prediction Models

Panel A: CEO Sample

	NT	IRAI	IR	AAER
Modified Jones Model Discretionary Accruals (DA)				
Sample Composition				
Total firm-quarters	9,593	9,593	9,593	9,593
Deceptive firm-quarters	1,566	1,096	951	203
Deceptive firm-quarters(%)	14.81	10.31	8.66	1.75
Area Under ROC Curve in %				
Linguistic-based model	54.83(5.66)	57.46(7.28)	59.24(8.77)	67.18(8.45)
DA	53.80(5.00)	53.68(3.58)	54.58(4.62)	62.98(5.94)
Corrected *t*-test for the difference	0.91	2.74^{***}	3.21^{***}	1.25
Performance-Matched Discretionary Accruals (PMDA)				
Sample Composition				
Total firm-quarters	9,220	9,220	9,220	9,220
Deceptive firm-quarters	1,509	1,064	924	201
Deceptive firm-quarters(%)	14.81	10.31	8.66	1.75
Area Under ROC Curve in %				
Linguistic-based model	55.32(6.89)	57.69(8.73)	59.05(9.94)	67.45(8.84)
PMDA	54.46(4.70)	55.43(5.91)	55.94(6.20)	62.96(6.60)
Corrected *t*-test for the difference	0.74	1.90^{*}	2.48^{**}	1.66
Audit Integrity Accounting Score (AI)				
Sample Composition				
Total firm-quarters	8,647	8,647	8,647	8,647
Deceptive firm-quarters	1,406	963	814	166
Deceptive firm-quarters(%)	14.81	10.31	8.66	1.75
Area Under ROC Curve in %				
Linguistic-based model	55.10(5.31)	57.58(7.16)	58.98(7.94)	68.40(8.11)
AI	56.01(6.43)	58.53(8.88)	60.12(9.89)	63.11(6.62)
Corrected *t*-test for the difference	−0.68	−0.74	−0.78	1.73^{*}
Dechow, Ge, Larson, and Sloan [2010] Model 1 (DGLSM1)				
Sample Composition				
Total firm-quarters	10,079	10,079	10,079	10,079
Deceptive firm-quarters	1,606	1,145	982	206
Deceptive firm-quarters(%)	14.81	10.31	8.66	1.75
Area Under ROC Curve in %				
Linguistic-based model	55.27(6.97)	57.67(7.95)	58.86(8.60)	67.31(9.88)
DGLSM1	52.71(3.39)	52.87(3.57)	53.65(3.51)	70.11(12.42)
Corrected *t*-test for the difference	2.21^{**}	4.25^{***}	3.72^{***}	−1.27
Beneish(1999) Model (BM)				
Sample Composition				
Total firm-quarters	7,561	7,561	7,561	7,561
Deceptive firm-quarters	1,236	879	750	130
Deceptive firm-quarters(%)	14.81	10.31	8.66	1.75
Area Under ROC Curve in %				
Linguistic-based model	54.88(4.84)	58.30(7.11)	59.68(8.97)	63.06(5.12)
BM	51.42(1.52)	52.40(2.22)	51.59(1.30)	60.84(4.78)
Corrected *t*-test for the difference	2.64^{***}	3.54^{***}	4.93^{***}	0.58

(Continued)

The only case where the financial-variables-based model performs better than the linguistic-based model is for model 1 from Dechow et al. [2011] relative to the model from CFO narratives under the AAER criterion (65.67% versus 72.58%). This is perhaps the fairest comparison

TABLE 9—*Continued*

Panel B: CFO Sample	NT	IRAI	IR	AAER
Modified Jones Model Discretionary Accruals (DA)				
Sample Composition				
Total firm-quarters	8,968	8,968	8,968	8,968
Deceptive firm-quarters	1,516	1,051	909	209
Deceptive firm-quarters(%)	15.37	10.70	9.04	1.90
Area Under ROC Curve in %				
Linguistic-based model	56.94(8.61)	58.37(8.56)	58.21(7.61)	65.78(8.12)
DA	53.88(4.15)	52.85(2.80)	53.83(3.46)	62.52(5.67)
Corrected *t*-test for the difference	2.54**	3.87***	2.80***	1.12
Performance-Matched Discretionary Accruals (PMDA)				
Sample Composition				
Total firm-quarters	8,624	8,624	8,624	8,624
Deceptive firm-quarters	1,458	1,019	882	206
Deceptive firm-quarters(%)	15.37	10.70	9.04	1.90
Area Under ROC Curve in %				
Linguistic-based model	57.53(8.85)	58.91(10.54)	58.66(8.77)	65.09(7.42)
PMDA	54.88(5.77)	54.62(4.54)	55.88(5.98)	62.36(6.69)
Corrected *t*-test for the difference	2.17**	3.66***	1.87*	1.02
Audit Integrity Accounting Score (AI)				
Sample Composition				
Total firm-quarters	8,049	8,049	8,049	8,049
Deceptive firm-quarters	1,364	926	790	165
Deceptive firm-quarters(%)	15.37	10.70	9.04	1.90
Area Under ROC Curve in %				
Linguistic-based model	56.82(6.98)	57.95(8.34)	58.15(8.09)	65.77(7.27)
AI	54.62(5.55)	56.86(6.64)	58.62(7.68)	62.16(5.94)
Corrected *t*-test for the difference	1.62	0.86	−0.31	1.14
Dechow, Ge, Larson, and Sloan [2010] Model 1 (DGLSM1)				
Sample Composition				
Total firm-quarters	9,568	9,568	9,568	9,568
Deceptive firm-quarters	1,573	1,122	957	217
Deceptive firm-quarters(%)	15.37	10.70	9.04	1.90
Area Under ROC Curve in %				
Linguistic-based model	56.90(8.74)	57.75(8.74)	58.37(8.86)	65.67(8.56)
DGLSM1	52.83(3.56)	53.35(3.25)	54.13(4.67)	72.58(13.82)
Corrected *t*-test for the difference	3.56***	3.28***	3.27***	−2.93***
Beneish [1999] Model (BM)				
Sample Composition				
Total firm-quarters	7,109	7,109	7,109	7,109
Deceptive firm-quarters	1,207	837	702	122
Deceptive firm-quarters(%)	15.37	10.70	9.04	1.90
Area Under ROC Curve in %				
Linguistic-based model	58.48(9.21)	58.65(8.11)	58.82(6.96)	61.28(3.65)
BM	51.61(1.54)	52.27(2.10)	51.54(1.33)	62.16(4.62)
Corrected *t*-test for the difference	5.17***	4.39***	4.45***	−0.21

This table presents the AUC (the area under ROC) in percentages for pairwise tests comparing the models that predict deceptive instances using verbal cues and five models that use only financial variables for the NT, IRAI, IR, and AAER deception criteria (defined in table 3). The results for CEO (CFO) are presented in panel A (panel B). The statistical tests are based on 10-fold cross-validation repeated 10 times, which provides us with 100 out-of-sample performance measures. In pairwise tests, linguistic-based and financial-variables-based models are estimated and tested on the same split of data. In parentheses, we report the corrected resampled *t*-statistic for the null hypothesis of the mean AUC being equal to 50% (e.g., Nadeau and Bengio [2003], Bouckaert and Frank [2004]). Here, *, **, and *** denote the corresponding results significant at 10%, 5%, and 1% significance level (two tailed test).

because this financial model was developed to predict AAERs. Finally, the commercial audit integrity accounting score exhibits statistically equivalent performance to the linguistic-based models for both CEOs and CFOs.

7.5 ASSOCIATION OF DECEPTION SCORES WITH FUTURE EXCESS RETURNS

Although the linguistic-based models have some ability to identify deceptive CEO and CFO narratives, it is also of interest to see whether this information is associated with future excess returns. In particular, we expect to observe negative returns in the months following the "deceptive" conference call if the market gradually learns about misreporting after the call. In addition, this test provides an estimate of the economic value of predicting executive deception for a representative firm. To examine this question, we compute the risk-adjusted returns that are produced by an equally weighted portfolio of stocks with the highest estimated deception scores. This computation is implemented using calendar time portfolio formation and standard risk adjustment using the four-factor Carhart [1997] model that includes three Fama–French factors (Fama and French [1993]) plus the momentum factor of Carhart [1997].

For every calendar quarter, we estimate (using all data available prior to that quarter) the linguistic model, within-sample probabilities for each firm, and a cutoff corresponding to the top 5%, top 10%, and top 20% for the probability of deception. For example, using 2003 data, we obtain the coefficient estimates for the prediction model and the distribution of the predicted probability of deception. These estimates are then applied to conference calls during the quarter that includes January, February, and March of 2004. Firms with predicted deception probabilities above the cutoffs estimated using 2003 data are included in the portfolio starting in the calendar month after their conference call (i.e., a firm with a conference call in January will be in the portfolio starting in February). We assume that a firm added to the portfolio is held for a fixed period of three months (i.e., February, March, and April for our example). The linguistic models and the probability cutoffs are updated quarterly. Because the ending date for our conference calls is May 2007, our analysis produces 43 monthly portfolio returns from February 2004 to August 2007.[22]

The annualized alphas (or annualized average monthly excess returns unexplained by the four-factor Carhart [1997] model) associated with this trading strategy are reported in table 10.[23] We find that alpha estimates for

[22] We acknowledge that the trading strategy that we are using is not implementable because it requires knowledge about the revelation of restatements in future periods in order to estimate the deception model.

[23] Consistent with prior research, we also require that the monthly portfolios have at least 10 individual firms and drop months with less than 10 firms in the portfolio from the return computations. This restriction is imposed in order to mitigate the impact of heteroskedasticity on the statistical tests.

TABLE 10

Annualized Excess Returns Associated with a Trading Strategy Based on the Deception Score from the Linguistic Prediction Model

	NT	IRAI	IR	AAER
Panel A: CEO Sample				
top 5%	0.00	0.05	0.05	−0.00
	(0.03)	(0.03)	(0.03)	(0.02)
	[42]	[43]	[43]	[43]
top 10%	0.02	0.03	0.03	0.01
	(0.02)	(0.02)	(0.02)	(0.02)
	[43]	[43]	[43]	[43]
top 20%	0.01	0.01	0.01	0.01
	(0.02)	(0.02)	(0.02)	(0.01)
	[43]	[43]	[43]	[43]
Panel B: CFO Sample				
top 5%	-0.11^{***}	-0.07^{**}	-0.07^{**}	−0.03
	(0.03)	(0.03)	(0.03)	(0.03)
	[42]	[42]	[42]	[42]
top 10%	-0.09^{***}	-0.05^{*}	-0.05^{**}	-0.05^{*}
	(0.03)	(0.03)	(0.02)	(0.03)
	[43]	[43]	[43]	[43]
top 20%	-0.08^{***}	-0.07^{**}	-0.04^{**}	−0.03
	(0.02)	(0.03)	(0.02)	(0.02)
	[43]	[43]	[43]	[43]

This table summarizes the estimated annualized intercepts and standard errors (in parentheses) from the four-factor Carhart [1997] model:

$$R_{it} - R_{ft} = \alpha_i + \beta_{1i}(R_{Mt} - R_{ft}) + \beta_{2i} SMB_t + \beta_{3i} HML_t + \beta_{4i} MOM_t + \epsilon_{it},$$

where R_{it} is the equally weighted monthly return on the portfolio consisting of firms that have a quarterly conference call with a predicted probability of being deceptive in the top 5%, top 10%, or top 20%; R_{ft} is the risk-free rate (the one-month U.S. Treasury bill rate); R_{Mt} is the market return (the return on a value-weighted portfolio of NYSE, Amex, and NASDAQ stocks); SMB_t and HML_t are the size and value-growth returns of Fama and French [1993], MOM_t is the momentum return Carhart [1997]; and α_i is the unexplained monthly average excess return. Portfolio formation is done in the calendar time and covers the time period from February 2004 to August 2007. For every calendar quarter, we estimate the linguistic prediction model and compute the within sample percentile cutoffs for the probability of deception using all data available prior to that quarter. Using the estimated model and the percentile cutoffs, firms are sorted into n, $n \in \{20, 10, 5\}$ portfolios for the next calendar quarter. Firms are held in the portfolio for three months. To mitigate the impact of individual stocks, we require that the portfolio has at least 10 stocks. The number of months used for estimation of alpha is reported in squared brackets. *, **, and *** denote correspondingly annualized intercepts significant at 10%, 5%, and 1% significance level (two-tailed test, critical values are from t-distribution).

the portfolios based on the CEO narratives are insignificant for all cutoff-deception criteria combinations. However, the annualized alpha estimates for the CFO model range from −4% to −11% depending on the criterion and the probability cutoff. It is interesting to note that the excess returns for the AAER model are the smallest in table 10. This is an unexpected result because AAERs are the most restrictive form of misreporting. We believe that this result occurs because AAERs are somewhat rare events relative to other forms of deception, and that our somewhat coarse trading strategy using the top 5%, top 10%, and top 20% for the probability of deception may not be appropriate for AAERs. Overall, at least for the linguistic-based

CFO model, the prediction of deceptive conference calls is economically valuable.[24]

7.6 EXTENSIONS

7.6.1. Combined CEO and CFO Models. The results in table 6 estimate separate models for the CEO and CFO. However, an obvious alternative approach is to simultaneously include linguistic variables for both executives in the prediction model. We find (untabulated) that the AUC is statistically higher than the AUC measures for individual CEO and CFO models, on average, by approximately 3%. Another way to combine the two executives is to use data for only the executive who speaks the most during the conference call. As noted above, CEO narratives have the most words in about 70% of the observations. We find (untabulated) that the AUC for this approach is statistically higher than the AUC for the CEO only model by slightly above 1% for all labels except AAER. However, for the CFO only model, the AUC for this approach is statistically lower by approximately 4% for the AAER label and statistically equivalent for the other labels. At this point, the best way to combine the linguistic responses for different individuals is an open methodological question. These preliminary results indicate that this choice can have an important impact on the predictive performance of the linguistic models.

7.6.2. Including All Word Categories from LIWC. As discussed in section 3, our selection of linguistic features was based on various contemporary theories about the word patterns used by deceptive individuals. Although many of the LIWC categories are used in our analysis, we dropped various features that did not seem to follow from the deception theories. This was done in order to avoid simply engaging in complete data mining. When we use all of the LIWC features, we find (untabulated) that the AUC is statistically higher by 3–6% than our more limited model, depending on the label used for accounting restatements. The new features that are statistically significant for both the CEO and CFO include, among others, causation (e.g., because, effect, hence, etc.) and inclusion (e.g., and, with, include) words that may be rationalized by the cognitive effort perspective. At the same time, some LIWC features that should be expected to be irrelevant (such as family, friends, home, etc.) are not statistically significant. Holding the data mining critique aside, these results indicate that our feature selection process missed some potentially important constructs. This is an expected

[24] We also computed the risk-adjusted returns using a value-weighted portfolio. Unlike the equally weighted results in table 10, we do not observe statistically significant returns for the value-weighted portfolio (i.e., the results in table 10 appear to be driven by small firms). Although certainly a speculation, it may be the case that conference calls provide more information about the value of small firms with less market attention and analyst coverage than large firms. Thus, we might expect deceptive conference calls to be associated with future results for small firms, but not larger firms.

outcome for initial research in a new area, and suggests an important avenue for subsequent studies.

7.6.3. CEOs with Accounting Background. One potential explanation for the differences in linguistic cues related to deception that we find for the CEO and CFO is that the language pattern might be affected by their professional training and experience. If their professional training affects the choice of linguistic cues, we would expect the CEOs with an accounting background to exhibit similar linguistic cues to CFOs. To examine this conjecture, we collect data on CEOs' professional background from BoardEx, which provides employment histories and education for individual executives. For our sample, in about 12% of instances, the CEO has an accounting background (measured by a CEO having a CPA or prior work experience as an auditor or as a CFO before becoming a CEO in the firm). In terms of specific tests, we include an indicator variable for a CEO having an accounting background and interact this variable with word categories. We do not find that CEOs with accounting training exhibit linguistic patterns similar to those of CFOs. Thus, the differences that we find between CEOs and CFOs do not appear to be explained simply by their professional training.

7.6.4. Incentives to Misreport. A maintained assumption in this paper is that the executives know about misreporting at the time of the conference call (i.e., misreporting is intentional). The intentional nature of misreporting may be impossible to verify. However, it is possible to provide some corroborating evidence using measures of the manager's incentives to misreport. Specifically, the classification performance of the linguistic-based model should be stronger if an executive has a greater personal incentive to misreport because this behavior is likely to be intentional. Consistent with a number of prior studies (see Armstrong, Jagolinzer, and Larcker [2010] for a review), higher equity incentives may be associated with greater incentives to misreport. We rely on this incentives argument and estimate the linguistic-based models in the low, medium, and high terciles of equity incentives delta. We assume that the executive holds the same equity portfolio throughout the year as at the beginning of the year. The composition of the equity portfolio is recovered following Core and Guay's [2002] assumptions. For every quarter, we use end-of-the-quarter stock price and stock return volatility to compute equity portfolio delta. We find that the out-of-sample classification performance improves across equity delta terciles with a gain in the AUC of 2–6% for CEOs. The improvement is smaller, however, and more monotonic for the sample of CFOs. This preliminary finding is consistent with the model exhibiting better classification performance when executive incentives to misreport are the greatest.

7.6.5. Individual Fixed Effects. There is the possibility of spurious classification results based on the composition of our sample and the way our cross validation is performed. The source of this potential problem is that deception affects multiple conference calls for the same firm. At every run of the 10-fold cross-validation procedure, we split the sample randomly, and there is the possibility of several instances of deception for the same executive in both the estimation and test samples. If the style of communication and word choice is individual-specific and persistent over time, some of the classification accuracy in the testing sample may come from the fact that we have the same deceptive individual in the estimation sample. That is, an individual-specific fixed effect is producing performance, as opposed to a pattern of deceptive language. However, it is important to note that, because we are only fitting a single model using the word categories, any correlation between the estimation and test samples can either improve or hurt the performance depending on the similarity of deceptive responses across individuals.

We have also assumed that there is a common benchmark for truthful and deceptive language across all individuals in our sample. It is perhaps more reasonable to assume that each individual has his/her own mode for truthful and deceptive language. This suggests that we should be able to improve the classification performance by adjusting simple word counts for individual fixed effects. Although this is a reasonable approach, it will be difficult to develop a good estimate for a normal (truthful) word count benchmark for an individual because our time series is limited. If the sampling variation for the individual fixed effect is large (due to a small number of time-series observations), the power of our tests with individual fixed effects will be substantially reduced.

We estimate a new set of results where word counts are adjusted by the average for an individual over all previous quarters (requiring a minimum of two quarters and excluding deceptive instances). Although the constraints from the fixed effects methodology reduce the sample size, we continue to find a significant classification performance in the linguistic models. The logistic models that use the adjusted word categories have an AUC that is statistically greater than the AUC for a random classifier by 6–14% for the sample of CEOs and by 6–12% for the sample of CFOs (untabulated). The adjusted models have statistically equivalent classification performance as unadjusted models except for CEOs where, under the NT criterion, the predictive performance of the adjusted model deteriorates (63.60% for the unadjusted model versus 56.39% for the adjusted model). These results suggest that our classification results using the unadjusted word categories are not likely to be entirely spurious.

The significance of the variables for adjusted word categories models does substantially change relative to the results from the unadjusted models (untabulated). The most significant change is that the reference to the general knowledge category is not statistically associated with deception for the adjusted analysis. However, as theorized, the self-reference category for CEOs has a statistically negative association with deception. If the number of adjusted self-references increases by 29, the odds of deception decrease by a factor of 0.54 for the NT criterion to 0.41 for the IR criterion for CEOs. Another important change for CEOs is that the negation words now have a significant positive association with deception. If the number of negations increases by 29, the odds of deception increase by a factor of 3.02 for the IRAI and 4.22 for the IR criterion. For CEOs, extreme positive emotion words are still strong predictors of deception and the effect of an increase in this category by 29 words increases the odds of deception by 2.47 for the IRAI and 3.07 for the IR label. For CFOs, the results after adjustment become considerably weaker. Impersonal pronouns and certainty words exhibit a positive association with the likelihood of deception and nonextreme positive emotion words have a negative association with deception.[25]

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 narratives 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 restated in subsequent time periods. Because the CEO and CFO are likely to know that financial statements have been manipulated, we are able to reasonably identify which executive discussions are actually "deceptive." Thus, we can estimate a linguistic-based model for detecting deception and test the *out-of-sample* performance of this classification method.

Our linguistic classification models based on CEO or CFO narratives perform significantly better than a random guess by 6–16%. In terms of linguistic features of the narratives, both CEOs and CFOs use more references to general knowledge, fewer nonextreme positive emotion words, and fewer shareholder value references. However, the pattern of deception for CEOs differs from that for CFOs. Specifically, CEOs use more extreme positive

[25] We also computed linguistic measures using the difference in the cues between the MD and Q&A, The resulting classification models were statistically better than random for the CEO, but not for the CFO. An important issue for future research is how (or perhaps whether) to adjust the linguistic measures for this predictive task.

emotion words and fewer anxiety words. In contrast, CFOs use more negation words and, for the most restrictive deception criterion (AAER), they use more extreme negative emotion words and swear words. In addition, under less restrictive criteria, deceptive narratives of CFOs contain fewer self-references and fewer impersonal pronouns.

In terms of predictive performance, linguistic-based models either dominate or are statistically equivalent to five contemporary models that are based on the accounting and financial variables. Finally, a trading strategy for the representative firm based on the CFO linguistic model produces a statistically significant annualized alpha (estimated using four-factor Carhart [1997] model) between −4% and −11%, depending on the deception criterion and portfolio selection method. The results for the CEO linguistic model do not produce a statistically significant alpha. Based on the strength of these exploratory performance results, we believe 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 rely on a general psychosocial dictionary, LIWC, which may not be completely appropriate for capturing business communication. Fourth, 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 whole population of firms. Finally, our sample only covers the time period from September 2003 to May 2007. Because this is shortly after the implementation of Sarbanes–Oxley and many restatements were observed during this period, our results may not generalize to time periods with fewer regulatory changes.

In terms of future research, it would be useful to refine general categories to business communication. It would also be desirable to adapt natural language processing approaches to capture the *context* of word usage and the choice of phrases for identifying deceptive executive behaviors. We have considered the conference call as a whole in our statistical analyses. However, it may be the case that there are better verbal cues for identifying deception in answers to questions related to specific accounts that were actually manipulated. It might be important to assess whether a question in the Q&A has an overall aggressive or friendly tone. This approach might enable us to refine the word categories or develop an alternative weighting scheme for computing a deception index.

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 (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): 1259–94.

ARMSTRONG, C. S.; A. D. JAGOLINZER; AND D. F. LARCKER. "Chief Executive Officer Equity Incentives and Accounting Irregularities." *Journal of Accounting Research* 48 (2010): 225–71.

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* 1 (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.

BENEISH, M. D. "The Detection of Earnings Manipulation." *Financial Analysts Journal* 55 (1999): 24–36.

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 (2005): 313–29.

BOUCKAERT, R. R., AND E. FRANK. *Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms*, in *PAKDD*. Berlin, Heidelberg: Springer-Verlag, 2004: 3–12.

CAMERON, A. C.; J. B. GELBACH; AND D. L. MILLER. "Robust Inference with Multi-Way Clustering." Working Paper 327, National Bureau of Economic Research, 2006.

CARHART, M. M. "On Persistence in Mutual Fund Performance." *The Journal of Finance* 52 (1997): 57–82.

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." Working paper, London Business School. 2009. SSRN eLibrary, Web site, http://ssrn.com/paper=1458478.

COURTIS, J. K. "Corporate Report Obfuscation: Artefact or Phenomenon?" *The British Accounting Review* 36 (2004): 291–312.

DAHL, D. B. xtable: Export Tables to LaTeX or HTML, 2009.

DAINES, R. M.; I. D. GOW; AND D. F. LARCKER. "Rating the Ratings: How Good Are Commercial Governance Ratings?" *Journal of Financial Economics* 98 (2010): 439–61.

DAS, S. R., AND M. Y. CHEN. "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web." *Management Science* 53 (2007): 1375–88.

DAVIS, A. K.; J. M. PIGER; AND L. M. SEDOR. "Beyond the Numbers: Managers' Use of Optimistic and Pessimistic Tone in Earnings Press Releases." Working paper, University of Oregon and DePaul University, 2007. SSRN eLibrary, Web site, http://ssrn.com/paper=875399.

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* 28 (2011): 17–82.

DECHOW, P. M.; R. G. SLOAN; AND A. P. SWEENEY. "Detecting Earnings Management. *The Accounting Review* 70 (1995): 193–225.

DEMERS, E. A., AND C. VEGA. "Soft Information in Earnings Announcements: News or Noise?" Working paper, INSEAD and Board of Governors of the Federal Reserve System, 2010. SSRN eLibrary, Web site, 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 (2003): 74–118.

DIETTERICH, T. G. "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms." *Neural Computation* 10 (1998): 1895–923.

EFRON, B., AND R. J. TIBSHIRANI. *An Introduction to the Bootstrap*. Boca Raton, London, New York, Washington D.C.: Chapman & Hall/CRC, 1994.

FAMA, E. F., AND K. R. FRENCH. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1993): 3–56.

FAWCETT, T. "An Introduction to ROC Analysis." *Pattern Recognition Letters* 27 (2006): 861–74.

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): 1–54.

FRIEDMAN, J.; T. HASTIE; AND R. TIBSHIRANI. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (2009): 1–22.

GOW, I. D.; G. ORMAZABAL; AND D. J. TAYLOR. "Correcting for Cross-Sectional and Time-Series Dependence in Accounting Research." *The Accounting Review* 85 (2010): 483–512.

HASTIE, T.; R. TIBSHIRANI; AND J. FRIEDMAN. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Corrected edition. New York: Springer-Verlag, 2003.

HENNES, K.; A. LEONE; AND B. MILLER. "The Importance of Distinguishing Errors from Irregularities in Restatement Research: The Case of Restatements and CEO/CFO Turnover." *The Accounting Review* 83 (2008): 1487–519.

HENRY, E., AND A. J. LEONE. "Measuring Qualitative Information in Capital Markets Research." Working paper, University of Miami, 2009. SSRN eLibrary, Web site, http://ssrn.com/paper=1470807.

HOBSON, J. L.; W. J. MAYEW; AND M. VENKATACHALAM. "Analyzing Speech to Detect Financial Misreporting." *Journal of Accounting Research* 50 (2012): 349–392.

HRIBAR, P., AND D. W. COLLINS. "Errors in Estimating Accruals: Implications for Empirical Research." *Journal of Accounting Research* 40 (2002): 105–34.

HUMPHERYS, S. L.; K. C. MOFFITT; M. B. BURNS; J. K. BURGOON; AND W. F. FELIX. "Identification of Fraudulent Financial Statements Using Linguistic Credibility Analysis." *Decision Support Systems* 50 (2011): 585–94.

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

JONES, K. L.; G. V. KRISHNAN; AND K. D. MELENDREZ. "Do Models of Discretionary Accruals Detect Actual Cases of Fraudulent and Restated Earnings? An Empirical Analysis." *Contemporary Accounting Research* 25 (2008): 499–531.

KNAPP, M. L.; R. P. HART; AND H. S. DENNIS. "An Exploration of Deception as a Communication Construct." *Human Communication Research* 1 (1974): 15–29.

KOTHARI, S.; A. J. LEONE; AND C. E. WASLEY. "Performance Matched Discretionary Accrual Measures." *Journal of Accounting and Economics* 39 (2005): 163–97.

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." *The Accounting Review* 84 (2009): 1639–70.

LI, F. "Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports?" Working paper, University of Michigan, 2006. SSRN eLibrary, Web site, http://ssrn.com/paper=898181.

LI, F. "Annual Report Readability, Current Earnings, and Earnings Persistence." *Journal of Accounting and Economics* 45 (2008): 221–47.

LI, F. "The Information Content of Forward-Looking Statements in Corporate Filings: A Naive Bayesian Machine Learning Approach." *Journal of Accounting Research* 48 (2010): 1049–102.

LOUGHRAN, T., AND B. MCDONALD. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66 (2011): 35–65.

LOUGHRAN, T.; B. MCDONALD; AND H. YUN. "A Wolf in Sheep's 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): 313–45.

NADEAU, C., AND Y. BENGIO. "Inference for the Generalization Error." *Machine Learning* 52 (2003): 239–81.

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): 665–75.

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. Austin, TX: LIWC.net., 2007. Available at http://homepage.psy.utexas.edu/homepage/Faculty/Pennebaker/Reprints/index.htm.

PETERSEN, M. A. "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches." *Review of Financial Studies* 22 (2009): 435–80.

PLUMLEE, M. A., AND T. L. YOHN. "Restatements: Investor Response and Firm Reporting Choices." Working paper, University of Utah and Indiana University, 2008. SSRN eLibrary, Web site, http://ssrn.com/paper=1186254.

PRICE, R. A.; N. Y. SHARP; AND D. A. WOOD. "Detecting and Predicting Accounting Irregularities: A Comparison of Commercial and Academic Risk Measures." *Accounting Horizons* (2011): Forthcoming. Available at http://ssrn.com/abstact=1912569.

R DEVELOPMENT CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2005.

RICHARDSON, S. A.; R. G. SLOAN; M. T. SOLIMAN; AND I. TUNA. "Accrual Reliability, Earnings Persistence and Stock Prices." *Journal of Accounting and Economics* 39 (2005): 437–85.

SCHOLZ, S. *The Changing Nature and Consequences of Public Company Financial Restaements 1997-2006*. Washington, D.C.: The U.S. Department of the Treasury, 2008.

SING, T.; O. SANDER; N. BEERENWINKEL; AND T. LENGAUER. "ROCR: Visualizing Classifier Performance in R." *Bioinformatics* 15 (2005): 3940–41.

TETLOCK, P. C.. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *Journal of Finance* 62 (2007): 1139–68.

TETLOCK, P. C.; M. SAAR-TSECHANSKY; AND S. MACSKASSY. "More Than Words: Quantifying Language to Measure Firms' Fundamentals." *Journal of Finance* 63 (2008): 1437–67.

THOMPSON, S. B. "Simple Formulas for Standard Errors That Cluster by Both Firm and Time." *Journal of Financial Economics* 99 (2011): 1–10.

TURNER, L., AND T. WEIRICH. "A Closer Look at Financial Statement Restatements: Analysing the Reasons Behind the Trend." *The CPA Journal* (December 2006): 13–23.

VRIJ, A. *Detecting Lies and Deceit: Pitfalls and Opportunities*, Second edition. Chichester, UK: John Wiley and Sons, 2008.

WITTEN, I. H., AND E. FRANK. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems, Second edition. Morgan Kaufmann, 2005.