Do Fraudulent Firms Produce Abnormal Disclosure?

Gerard Hoberg[*] and Craig Lewis[*]

April 20, 2015

ABSTRACT

Using text-based analysis of 10-K MD&A disclosures, we find that fraudulent firms produce verbal disclosure that is abnormal relative to strong counterfactuals. This abnormal text predicts fraud out of sample, has a verbal factor structure, and can be interpreted to reveal likely mechanisms that surround fraudulent behavior. Using a conservative difference-based approach, we find evidence that fraudulent managers grandstand good performance and disclose fewer details explaining the sources of the firm's performance. We also find new interpretable verbal support for existing hypotheses suggested in the literature, for example, that some managers commit fraud in order to improve their odds of raising capital at low cost.

[*] University of Southern California and Vanderbilt University, respectively. We thank Ken Ahern, Christopher Ball, Radha Gopalan, Kathleen Hanley, Yuan Li, Tim Loughran, Vojislav Maksimovic, Bill McDonald, Gordon Phillips, and Harvey Westbrook for excellent comments and suggestions. We also thank conference participants at the AFA meetings in Boston, and seminar participants at Columbia University, Duke University, George Washington University, London Business School, London School of Economics, U. S. Securities and Exchange Commission, University of Memphis, University of North Carolina at Chapel Hill, University of Notre Dame, University of Southern California, University of Tennessee, Knoxville, University of Washington, and U.S. Department of Treasury, Office of Financial Research. Any remaining errors are ours alone.

Many studies suggest that managers committing fraud likely do so to achieve various objectives such as getting access to low cost capital (Dechow, Sloan and Sweeney (1996), Povel, Singh and Winton (2007) and Wang, Winton and Yu (2010)) or to conceal diminishing performance (Dechow, Ge, Larson and Sloan (2011)).[1] We examine the question of whether a firm's 10-K MD&A disclosure to the U. S. Securities and Exchange Commission (SEC) reflects the fact that a firm committed fraud. This issue should be particularly salient to managers committing fraud because the SEC is tasked with identifying and pursuing accounting, auditing and enforcement actions (AAERs) against such firms. These disclosures also are simultaneously submitted to the public, which uses them to inform its decisions to allocate capital among firms seeking financing and other resources.

The high degree of discretion associated with managements' discussion and analysis of its operating performance in the 10-K (MD&A) creates opposing trade offs regarding how a fraudulent firm might explain its results. On one hand, fear of detection might lead managers to disclose in a way that attempts to minimize detection. For example, a manager might under-disclose detailed explanations of their fraudulent revenue or expense calculations. On the other hand, fraudulent managers have incentives to abnormally disclose information that can further maximize the objectives that led them to commit fraud in the first place. For example, a manager who manipulated revenues to attract more equity investors might have incentives to grandstand the firm's revenue growth to further draw more investor attention.

We first examine whether fraudulent firms have verbal disclosure that is abnormal relative to two benchmarks. The first examines abnormal disclosure in the cross section that is common to fraudulent firms after controlling for the disclosure of industry peers of similar size and age. The second focuses on the time series behavior of the fraudulent firm itself, and examines if firms have disclosures that differ from their own disclosures in the years prior to and after their alleged fraud. We find strong and uniform support

[1] Dechow, Ge and Schrand (2010) provide a detailed review of fraud literature, and we summarize this literature in detail in Section I of this paper.

for our hypothesis that firms committing fraud have a strong common component in their disclosures, and this component is less prevalent among firms not committing fraud. This finding cannot be explained by the disclosure of industry peers or firm fixed effects, and this abnormal disclosure only appears during the specific years firms are committing fraud.

Before we address the question of why fraudulent firms have a common signature in their MD&A disclosure, we note that the possibility that MD&A might contain a signal that predicts whether firms have engaged in accounting fraud is an important research question in its own right. While it is further important to then test competing explanations, the simple ability to improve the prediction of accounting fraud is practically relevant to future researchers, investors and regulators alike. It also provides a stronger empirical foundation for future theoretical and empirical researchers to further examine hypotheses.

To understand why fraudulent firms have common abnormal disclosures, we first consider three hypotheses derived from the incentives of fraudulent firms to alter their verbal disclosures. These hypotheses specifically relate to MD&A text and are as follows: (1) fraudulent managers conceal details that might expose their fraudulent accounting, (2) fraudulent managers grandstand their good performance, and (3) fraudulent managers disassociate themselves from the fraudulent disclosure by avoiding references to themselves in MD&A. To test these three hypotheses, we note that a researcher needs interpretable text analytic methods, as these hypotheses are silent regarding predictions about quantitative accounting data.

To test these hypotheses, we consider content analysis using Latent Dirichlet Allocation (LDA) to identify interpretable verbal topics that firms involved in AAERs use more compared to peers not involved in AAERs. LDA is a topic modeling technique developed by Blei, Ng and Jordan (2003). It is a generative model solved using likelihood analysis that discovers clusters of text (referred to as "topics") that frequently appear in various documents. LDA is intuitively akin to a sophisticated text-based analog of widely used numerical factor analysis. We discuss LDA in greater detail in Section 5. Specifically, we consider the MD&A-based LDA factors from Ball, Hoberg and Maksimovic (2013), and examine which particular topics are associated with firms committing fraud. We supplement these existing quantitative methods by further extracting "representative paragraphs" from the corpus to more concretely associate each topic with a specific interpretation. We note that these calculations are fully automated and thus the qualitative content generated by them is not subject to researcher prejudice.

We find support for all three hypotheses. Specifically, we find that firms committing fraud abnormally under-disclose details regarding why the firm experienced its observed level of performance. This conclusion is based on a specific topic identified by the LDA methodology. The representative paragraph associated with this topic states:

> *"The increase was due primarily to the increase in sales, the decreased percentage of the general and administrative expenses and the decrease in depreciation and amortization expense.... The decrease was due primarily to the increased costs of operating the company-owned restaurants."*

The fact that fraudulent firms make less use of this LDA topic related to providing quantitative details is direct evidence that fraudulent firms disclose fewer details explaining their performance.

By separately examining instances of revenue fraud and expense fraud, we are further able to test the hypothesis that fraudulent firms grandstand the manipulated performance itself. For example, we find that firms engaged in revenue fraud abnormally disclose more of a topic that highlights the firm's revenue growth. The representative paragraph associated with this topic, for example, starts with the statement

> *"Revenues increased by $29.9 million, or approximately 27.4%, to $139.1 million in 1997 from $109.2 million in 1996."*

Regarding firms engaged in expense fraud, they disclose abnormally high levels of a topic that associates with a representative paragraph stating

> *"Research and development expenses increased 20.7% to $6,006,000 in 1996, and increased as a percentage of net sales to 10.0% in 1996 from 6.1% in 1995. The increases in research and development expenses were primarily due to the expansion of the research and development staff, and expenses associated with its research and development facility."*

Because an active R&D platform is highly valued in the stock market, this is consistent with managers inflating the firm's perceived growth options by committing expense fraud that is reinforced by effervescent disclosure.

We also find evidence that managers tend to disassociate themselves from the fraudulent disclosure during years the firm is engaged in fraud. For example, fraudulent firms disclose abnormally low levels of an LDA topic that touts the manager's overall participation in the firm's vision and strategy. The representative paragraph from this topic begins with

> *"Since joining the company in January 1998, the new chief executive officer, along with the rest of the company's management team has been developing a broad operational and financial restructuring plan."*

Because LDA provides a full taxonomy categorizing information in the MD&A section of the 10-K as a whole (see Ball, Hoberg and Maksimovic (2013)), it is also easy for us to test two hypotheses from the existing literature based on the incentive to initiate fraud. We highlight these hypotheses because they have particularly strong predictions that relate to the content that typically appears in MD&A.

First, we consider the hypothesis that managers initiate fraud to reduce their cost of capital or to alleviate financial constraints (Dechow, Sloan and Sweeney (1996), Povel, Singh and Winton (2007), and Wang, Winton and Yu (2010)). Consistent with this hypothesis, we find that fraudulent firms under-disclose a topic associated with the firm's required summary of liquidity challenges. For example, the representative paragraph associated with this topic starts by saying

> *"The company believes that its current cash, cash equivalents and short-term investment balances and cash flow from operations, if any, will be sufficient to meet the company's working capital and capital expenditure requirements for at least the next twelve months. Thereafter, the company may require additional funds..."*.

Hence fraudulent managers disclose less content that might indicate the true limitations of their firm's supply of liquidity. Following Hoberg and Maksimovic (2015), this can indicate to investors that the firm is less constrained.

Given the popularity of this specific hypothesis (which links fraud to manipulating liquidity and the cost of capital) in the literature, we also consider a quasi-natural experiment based on exogenous forced mutual fund selling following Coval and Stafford (2007) and Edmans, Goldstein, and Jiang (2012). We find that when firms face exogenous negative liquidity shocks, they increase their disclosure of text that correlates highly with the abnormal text of fraudulent firms. We also find that the observed level of ex-post AAERs also increases. This suggests that firms increase their use of manipulative text and are more likely to commit fraud when their liquidity deteriorates.

Finally, we also find support for the hypothesis that fraudulent firms likely have motives that are linked to the pricing of acquisitions (Erickson and Wang (1999) and Wang (2013)), as we observe fraudulent firms disclosing abnormally high levels of a topic that discusses acquisitions.

The remainder of this article is organized as follows. Section I presents our hypotheses, Section II describes our data and methodology, and Section III presents our data and methods. Section IV presents our abnormal disclosure regressions and Section V presents content analysis. Section VI presents our quasi-natural experiment based on equity market liquidity, and Section VII concludes.

1 Hypotheses

Many studies examine the links between accounting variables and AAERs. For example, Feroz, Park, and Pastena (1991) and Karpoff, Lee and Martin (2008b) examine the issues that motivate fraud and their consequences. We refer readers to Dechow, Ge and Schrand (2010) for a thorough review.

A key premise of our hypotheses is that MD&A is an informative setting for understanding fraud. This premise relates both to understanding the incentives to commit fraud and to the possibility of improved fraud detection. To this end, we note that manipulations of revenues and expenses are the basis for the majority of AAER fraud allegations. In MD&A, managers discuss both as part of their required discussion of annual performance.

1.1 Hypotheses Based on Fraud Verbal Disclosure Incentives

We start with three hypotheses based on the incentives of fraudulent managers to alter their verbal disclosures. We label these hypotheses “H1A”, “H1B’, and “H1C”.

H1A [Managerial Detail Concealment]: *Fraudulent managers under-report details explaining the firm's performance.*

H1B [Managerial Grandstanding]: *Fraudulent managers grandstand the firm's strong growth and the quantities they manipulated to increase the impact of the manipulation.*

H1C [Managerial Disassociation]: *Fraudulent managers reduce the extent to which the management team itself is mentioned in the MD&A.*

The three hypotheses above are motivated by managerial incentives. Managers have incentives to conceal details associated with the firm's performance to increase the cost of detection. They also have incentives to grandstand because they committed fraud to make the firm look stronger than it is, and verbal grandstanding can further maximize that objective. This notion of grandstanding is related to the finding in Kedia and Philippon (2009) that fraudulent firms invest and hire more workers than they might optimally need. Our hypothesis predicts that managers will complement this form of real investment inflation with disclosure grandstanding to portray a unified signal to investors about the firm's prospects. Grandstanding might also be more likely when investors might have high expectations for future growth (Dechow, Ge, Larson and Sloan (2011) and Skinner and Sloan (2002)).

We also note that, although they initially seem at odds, H1A and H1B are not mutually exclusive. For example, a manager who commits revenue fraud might grandstand the fact that their firm experienced substantial growth (H1B). Yet, at the same time, they might also provide very little in the way of details explaining which specific aspects of their operations contributed most to the firm's performance (H1A). These hypotheses, which yield unique textual predictions for each type of fraud (revenue and expense fraud), motivate our empirical framework where we examine (A) all AAERs, (B) only AAERs that are based on revenue fraud and (C) only AAERs that are based on expense fraud.

H1C is also consistent with incentives. Managers are aware that the fraud might ultimately be caught. Hence they have an incentive to distance their own names and reputations from the disclosure, especially disclosure that discusses the fraudulent accounting itself. In the context of MD&A, we would thus expect fewer mentions of the management team itself when the firm is committing fraud as compared to materially similar firms that are not committing fraud.

1.2 Hypotheses Based on Fraud Initiation Incentives

We now focus on two hypotheses from the existing literature that make additional predictions regarding the content in the MD&A section of the 10-K. For example, MD&A covers discussions of performance, firm investment, capital structure, liquidity needs, and growth expectations.

H2A [Fraud to Reduce the Cost of Capital]: *Managers commit fraud to increase the likelihood that they will be able to successfully raise capital at a low cost. This hypothesis is discussed by Dechow, Sloan and Sweeney (1996), Povel, Singh and Winton (2007), and Wang, Winton and Yu (2010).*

H2B [Fraud to Reduce Cost of M&A]: *Managers commit fraud to achieve more favorable stock exchange ratios in stock-based M&A transactions. This hypothesis is discussed by Erickson and Wang (1999) and Wang (2013).*

Each hypothesis has direct predictions regarding the firm's verbal disclosure in its MD&A. H2A predicts that the firm will under-disclose information indicating liquidity problems[2] relative to similar firms not committing fraud. H2B predicts a higher incidence of discussions of acquisitions during fraud-years.

Our primary advantage in testing these hypotheses is that text allows us to examine specialized predictions that pertain to specific interpretable text in the 10-K. This can yield more transparent conclusions as we do not have to rely on proxy relationships between accounting variables. We also go further and consider a quasi-natural experiment to further test H2A, one of the most widely discussed motives for fraud. Here we consider exogenous forced mutual fund selling shocks following Coval and Stafford (2007) and Edmans, Goldstein, Jiang (2012). In particular, following the authors, we consider forced selling by non-sector-specific funds as a shock to the equity market liquidity of exposed firms. These results support H2A, as we find that exogenous shocks to equity market liquidity increase

[2] We focus on liquidity discussions because firms are required to disclose liquidity problems in MD&A (see for example Hoberg and Maksimovic (2015)). The authors also note a high degree of heterogeneity in this liquidity disclosure, which indicates that there is likely adequate power to test the current hypothesis.

the firm's use of text that associates most strongly with fraud.

We also note that although we are able to test many hypotheses using MD&A, we cannot test every hypothesis. For example, we believe that MD&A is unsuitable for testing the link between executive compensation motives and fraud (see for example Johnson, Ryan and Tian (2009), Goldman and Slezak (2006), and Burns and Kedia (2006)). The reason is that executive compensation is usually discussed in Item 11 of the 10-K, which is distinct from the MD&A (Item 7 of the 10-K). However, analogous tests based on other sections of the 10-K offer excellent potential for future research.

2 Data and Methodology

We create our sample and our key variables using two primary data sources: COMPUSTAT and the text in the Management's Discussion and Analysis section of annual firm 10-Ks (extracted using software provided by metaHeuristica LLC).

We first extract COMPUSTAT observations from 1997 to 2008 and apply a number of basic screens to ensure our examination covers firms that are non-trivial publicly traded firms in the given year. We start with a sample of 87,887 observations with positive sales, at least $1 million in assets, and non-missing operating income. We also discard firms with a missing SIC code or a SIC code in the range 6000 to 6999 to exclude financials, which have unique disclosures (especially because MD&A covers financial market liquidity and capital structure). This leaves us with 71,637 observations. After requiring that observations are in the CRSP database, we have 60,853 observations. Our sample begins in 1997 because this is the first year of full electronic coverage of 10-K filings in the Edgar database. Our sample ends in 2008 as this is the final year of our AAER database.

We also require that each observation has a machine readable MD&A section with a valid central index key (CIK) link to the Compustat database.[3] We use software provided by metaHeuristica to web crawl and to extract the MD&A section from each 10-K. MetaHeuristica uses natural language processing to parse and organize textual data, and its pipeline employs "Chained Context Discovery" (See Cimiano (2010) for details). The

[3] We use the WRDS SEC Analytics package to link 10-Ks to Compustat.

majority of 10-Ks (over 90%) have a machine readable MD&A section. The primary reason why a firm might not have a machine readable MD&A is when it is "incorporated by reference," and is not in the body of the 10-K itself.[4] These requirements leave us with a final sample of 49,039 firm-year observations having adequate data.

2.1 Accounting and Auditing Enforcement Releases

We obtain data on Accounting and Auditing Enforcement Releases (AAERs) from the Securities and Exchange Commission website[5]. Our hand collected sample includes AAERs indicating fraudulent behavior from 1997 to 2008. In addition to firm identifying data, which is needed to link AAER firms to our Compustat universe, we also collect the filing date of each AAER, and the beginning and ending dates each AAER alleges fraudulent activity. Our AAER dummy is set equal to one for firm fiscal years ending in calendar years that overlap with these begin and end dates. This is our primary variable of interest, and we focus on how disclosure varies during these AAER years. For an example of how it is calculated, consider a firm that has a June 30 fiscal year end and committed a fraud that began in July 2009 and ended May 2011. The AAER dummy variable would be set equal to one for fiscal year ends 2009, 2010, and 2011. If this same firm initiated the fraud on August 2009 instead, the AAER dummy variable would only be coded one for fiscal years 2010 and 2011.[6]

For each AAER, we also identify a year that is definitively prior to the alleged fraudulent activity, and a year that is definitively subsequent to the public release of the AAER by the SEC. We refer to these as the pre-AAER year and the post-AAER year. Our assessing disclosure in three critical periods (prior to, during, and after the alleged fraud) serves two purposes. First, this serves as a placebo test, as we expect a strong signal only during the years of fraudulent activity, and not in the years prior to or after the alleged fraud. Second, this allows us to understand the disclosure life cycle of fraudulent firms.

Due to the approximate nature of stated fraud periods, we take a conservative approach

[4] The typical scenario under which a MD&A section is incorporated by reference is when the annual report is submitted along with or referenced by the 10-K, and thus MD&A is not in the 10-K itself.

[5] http://www.sec.gov/divisions/enforce/friactions.shtml

[6] If a firm has more than one AAER, and the periods of alleged fraud overlap, we set the AAER dummy to one for any fiscal period where at least one fraud event is alleged.

when identifying the pre-AAER year and the post-AAER year. We define the pre-AAER year as the fiscal year preceding the first full calendar year that precedes the alleged fraud period. This ensures that, even with 10-K reporting delays and potential approximate identification of the fraudulent period, the pre-AAER year has disclosure that is unlikely to be contaminated by disclosure associated with the fraud. We identify the post-AAER year as the fiscal year end in the calendar year that is subsequent to the calendar year in which the AAER is announced to the public on the SEC website. This ensures that the firm had adequate time to update its disclosure subsequent to the alleged fraud.

2.2 Disclosure Industry Similarity

In this section, we focus on identifying the disclosure similarity between a firm and its size-age-industry matched peers. We refer to this as our "Industry Similarity" measure. Our approach of identifying common industry disclosure is related to Hanley and Hoberg (2010), who examine IPO pricing.

We first group all firms into bins based on industry (two-digit SIC codes), size and age. In particular, for each industry group in each year, we create a small firm and a large firm bin based on the median size of firms in each industry bin. We then divide bins once again based on median age (listing vintage). We thus have four bins for each SIC-2 industry, and each of the four bins has nearly the same number of firms. If a given bin has less than two firms, we exclude it from the rest of our analysis. Given that our two-digit SIC categories are rather coarse, this requirement affects less than one percent of our sample. We also note that our findings are robust to only using industry bins rather than these industry-size-age bins. We use these more refined bins because we expect material systematic differences in disclosure across firms of different size and age. We refer to a firm's peers in its industry, size, and age bin as its "ISA peers".[7]

Following standard practice in text analytics, we first discard stop-words and then convert the text in each firm's MD&A into vectors of common length across all firms. We define a "stop word" as any word appearing in more than 25% of all MD&A filings in

[7] In unreported results, we examine if our results are robust to further excluding fraudulent firms from the group of ISA peers. This has little influence on our results because fraudulent firms are relatively rare in our sample.

the first year of our sample (1997). The length of the vectors we create is based on the universe of remaining words. Because our calculations are computationally intensive, we restrict attention to words appearing in the MD&A of at least 100 firms in the first year of our sample (1997).[8] The resulting list of words is stable over time, as 99.1% of randomly drawn words using our 1997-based screen would be included using an analogous screen based on 2008. Each firm-year's MD&A is thus represented by its word distribution vector $W_{i,t}$. This vector sums to one, and each element indicates the relative frequency of the given word in the given MD&A. Our use of 1997 data to determine the word universe is meant to be conservative, as we avoid any look ahead bias in our later regressions that are based on an out of sample predictive framework.

To quantify disclosure similarity with ISA peers, we next compute the average word usage vector for a given firm's ISA peers excluding itself ($ISA_{i,t}$). It is important that this average excludes the firm itself, as skipping this step would create a mechanistic degree of similarity for firms in less populous bins. Our measure of industry disclosure similarity (H_{it}) is the cosine similarity between $W_{i,t}$ and $ISA_{i,t}$.

$$H_{i,t} = \frac{W_{i,t}}{\sqrt{(W_{i,t} \cdot W_{i,t})}} \cdot \frac{ISA_{i,t}}{\sqrt{(ISA_{i,t} \cdot ISA_{i,t})}} \tag{1}$$

The cosine similarity is a standard technique in computational linguistics (See Sebastiani (2002) for example). It is also easy to interpret, as two documents with no overlap have a similarity of zero, whereas two identical documents have a cosine similarity of 1. Finally, by virtue of its normalization of vectors to unit length, this method also has the good property that it correlates only modestly with document length.

2.3 Disclosure Fraud Similarity

In this section, we construct measures of the extent to which firms engaged in fraudulent behavior produce common disclosure, while controlling for the disclosure of ISA peers. We first compute abnormal disclosure for each firm ($AW_{i,t}$) as follows:

[8] This results in a vector length of roughly 10,000 words. We also note that our findings are robust to instead using a stricter screen based on 5,000 words. Because we also do not see a material degree of improvement in going from 5,000 to 10,000 words, we thus conclude that our universe is sufficiently refined to provide a relevant signal for testing our key hypotheses.

$$AW_{i,t} = W_{i,t} - ISA_{i,t} \tag{2}$$

We note that we only include non-fraudulent ISA peers in this calculation. The resulting vector sums to zero, as W_{it} and ISA_{it} each sum to one. We next compute the average deviation from industry peers made by firms known to be involved in SEC AAER enforcement actions (where N_{AAER} is the number of AAER firm-years from 1997 to 2001):

$$AAER_{vocab} = \sum_{j=1,\ldots,N_{AAER}} \frac{AW_j}{N_{AAER}} \tag{3}$$

Note that the vector $AAER_{vocab}$ does not have a time subscript, as we are summing the unique disclosures over all AAERs in a given universe. We note here that we only tabulate this average over firms with an AAER dummy of one in the years 1997 to 2001. We do not use the years 2002 to 2008 for training as we wish to preserve these years for assessing the out of sample performance of our fraud similarity variable in later tests. Our results are stronger if we instead use our entire sample for the computation of the $AAER_{vocab}$. Our approach ensures that results are not driven by look ahead bias. We then define the fraud profile similarity (we will also refer to this as the "fraud score") of a firm in a given year F_{it} as the cosine similarity between $AW_{i,t}$ and $AAER_{vocab}$ as follows (we also exclude firm i itself from the computation of $AAER_{vocab}$ to avoid any mechanistic correlations):

$$F_{i,t} = \frac{AW_{i,t}}{\sqrt{(AW_{i,t} \cdot AW_{i,t})}} \cdot \frac{AAER_{vocab}}{\sqrt{(AAER_{vocab} \cdot AAER_{vocab})}} \tag{4}$$

3 Data and Summary Statistics

Table 1 displays summary statistics for our panel of 49,039 firm-year observations from 1997 to 2008 having machine readable MD&As. 1.5% of firm year observations are AAER-years. As it is based on cosine similarities between positive and negative word vectors, the Fraud Similarity Score has a distribution in the interval [-1,+1] and a mean that is close to zero. Intuitively, because AAER years are rare, the average firm does not have a vocabulary that correlates highly with fraudulent firms. The industry similarity score is based on cosine similarities of non-negative vectors, and is bounded in the interval [0,1]. Its mean of 0.667 indicates that the average firm shares a substantial amount of disclosure with its ISA peers. However, the average firm also has much unique content.

[Insert Table 1 Here]

Table 2 displays Pearson correlation coefficients. The positive 8.2% correlation between the AAER dummy and the fraud similarity score (significant at the 1% level) foreshadows our later multivariate results. This suggests that firms involved in potentially fraudulent activity have abnormal disclosure relative to ISA peers that is common among AAER firms. The correlation between the AAER dummy and industry similarity is much weaker at 2.6%. Remarkably, the fraud similarity score is more correlated with the AAER dummy than any of the other displayed variables including firm size (7.0% correlation).

[Insert Table 2 Here]

Fraud similarity is 9.0% correlated with industry similarity (significant at the 1% level). Given that both variables are functions of firm disclosures, this is somewhat modest. The modest result is by construction, as fraud similarity is a function of abnormal disclosure after controlling for ISA peers. We also note that fraud similarity correlates little with firm size, which also relates to its construction based on size-adjusted peers (in addition to industry and age adjustments). These aspects of our variables help to ensure a clear interpretation in both univariate and multivariate settings. Finally, these modest correlations indicate that multicollinearity is unlikely to be a concern.

[Insert Table 3 Here]

Table 3 displays time series summary statistics regarding AAER-year observations in our sample from 1997 to 2008. The table shows a peak in 2000 to 2002 following the internet bubble's collapse, and also a steady stream of AAER years throughout our sample with the exception of the last three years, where the incidence rate is lower. As our analysis controls for both industry and time effects, as well as other controls, these features of our data cannot explain our results. We also note that, in all, 2.9% of our sample firms (249 of 8510) were involved in an AAER at some point in time in our sample. The relatively low rate of AAERs during the financial crisis of 2007 and 2008 does not necessarily point to a reduction in the rate of fraud but is more likely explained by a change in the SEC's priorities following the crisis.

3.1 Initial Evidence of Disclosure Differences

In this section, we explore the distributional features of our industry similarity and fraud similarity measures, and their links to observed AAER Enforcement actions. In Table 4, we sort firms into deciles based on their fraud similarity and industry similarity measures. We then report the fraction of firms in each decile that are involved in AAERs.

[Insert Table 4 Here]

Panel A of Table 4 displays these results for our entire sample, and shows that the incidence rate of AAERs is strongly positively correlated with the fraud similarity decile or in the industry similarity decile in which a firm resides. The results are economically large and decile sorting is close to monotonic. Regarding fraud similarity, the incidence rate of AAERs in decile 10 is 3.7% compared to just 0.5% for decile 1. The positive link between industry similarity and AAER incidence is weaker with high to low decile range of 2.7% to 1.0%.

Panel B of Table 4 displays analogous results for the out of sample period from 2002 to 2008. We remind readers that the key vocabulary used to compute the vocabulary associated with fraudulent firms is computed only using data from 1997 to 2001 (see Section 2.3). Hence, our assessment of the link between AAERs and the fraud scores in 2002 to 2008 is an out of sample test on all levels. We continue to observe strong positive associations with AAER incidence rates for fraud similarity, and the inter-decile range is 0.7% to 2.2%. Our later tests will show that our results for fraud similarity are especially strong both statistically and economically, and are also robust to multivariate regressions including controls for firm and industry fixed effects. In contrast, industry similarity plays a more passive role, and its correlation with AAERs is not robust to firm fixed effects. The results in this section indicate that the vocabulary used by AAER firms, that is distinct from industry-size-age peers, has remained stable over time.

3.2 Fraud Similarity Distributions

In this section, we examine the distribution of fraud similarity. Figure 1 shows the empirical density function of this variable over its domain [-1,1]. The distribution is centered near zero and is nearly bell shaped. However, it is somewhat asymmetric and right skewed, indicating that observations are potentially drawn from a mixed distribution where potentially fraudulent firms have a higher mean than non-fraudulent firms. The solid line shows the reflection of the distribution around the y-axis and illustrates the extent of the right skewness. As the figure indicates, the amount of probability mass that differs from the reflection is 2.55% of the total mass. This is materially larger than the observed 1.5% AAER rate indicated in Table 1.

[**Insert Figure 1 Here**]

We consider whether the rate of undetected fraud can be approximated. To do so, we make two assumptions that are unique to this exercise. These assumptions are not relevant to the tests in other parts of the paper and are employed here to provide intuition for the possible economic importance of undetected fraud. First, we assume that non-fraudulent firms have symmetrically distributed fraud similarities. Second, we assume that firms engaged in undetected fraud have a similar distribution compared to those that have been detected. This allows us to estimate the extent of undetected fraud based on how many firms would have to be removed from the sample to eliminate the observed asymmetry. We note that whether these assumptions hold likely depends critically on the nature of how fraud is detected, and whether the mechanism strongly relates to verbal text in the disclosure. Although it is unlikely that these assumptions hold precisely, the results in Dyck, Morse and Zingales (2010) suggest that they might only be weakly violated. In particular, the authors find that the primary consumers of 10-Ks (investors, the SEC, and auditors) play only a small role in detecting fraud. Employees and the media play a larger role.

[**Insert Figure 2 Here**]

We next assess the extent to which the removal of known AAER firm-years reduces asymmetry. Figure 2 plots the density function of fraud similarity separately for firms not involved in AAERs (upper figure) and involved in AAERs (lower figure). The figure shows that the density function retains a substantial degree of asymmetry even when known AAER firm-years are excluded, as the right-skewed mass only decreases from 2.55% to 2.10%. We thus compute the upper bound regarding the rate of undetected fraud as the fraction of the sample that would have to be removed to eliminate all observed asymmetric mass. This calculation suggests that just 17.6% ($\frac{2.55-2.10}{2.55}$) of fraudulent firms have been detected and hence fraud is 5.6x as pervasive as observed. We compute a lower bound by assuming that the 2.1% of remaining asymmetry in Figure 2 is due to 2.1% of undetected firms being engaged in fraud. This would imply that fraud is 2.4x as pervasive as observed. Because the observed rate of known AAER firm years is 1.5%, these estimates indicate that the actual rate of committed AAERs likely lies in the range (3.6%, 8.5%) of all firm-years. This range is substantially higher than the 1.5% detection rate in our sample.

The lower plot in Figure 2 further illustrates why our approach might have good power for estimating undetected fraud. The lower plot displays the density function of fraud similarity for firms that are known to be involved in AAERs. The figure shows a far higher degree of asymmetry than any of the other figures, indicating that fraud similarity is effective in separating AAER firms from non-AAER firms. The degree of asymmetric mass is 41.0%, which is far larger than the 2.1% in the upper figure.

Figure 3 displays fraud similarity scores over time: before, during and after a firm is involved in an AAER. We also explore the extent to which fraud similarity varies when a firm is involved in an AAER alleging a longer duration of fraud. In particular, we tag the three years that are prior to the calendar year in which the AAER indicates that the fraud began as the pre-fraud period, and the three years after the calendar year in which the AAER indicates that the fraud ended as the post-fraud period. We then consider up to three years of time during which an alleged fraud occurred. If a firm's alleged fraud period is three or more years, it will enter the average fraud similarity calculation for the first three of these years. If the firm's alleged fraud lasted only one or two years, it will only be included in the first and second fraud year calculations, respectively. To ensure robustness, we also consider this calculation only for firms that experienced a fraud period of at least three years.

[**Insert Figure 3 Here**]

The figure shows a trapezoidal pattern for fraud similarity. During the three years preceding the alleged fraud, the average fraud similarity slowly increases from nearly zero to 0.025. During the period of alleged fraud, this score more than doubles to over 0.05, and remains near this level during the years of alleged fraud. After the period of alleged fraud ends, fraud similarity then drops sharply to 0.025 and then dissipates to zero. Because the AAER is only announced after the fraud has occurred, these results provide strong time series evidence that we have identified a set of disclosure vocabularies that are used more by firms alleged to have committed fraud relative to those that have not. Because the figure reports scores for the same firms in all periods, these results are stark and automatically account for firm fixed effects.

4 Disclosure and Fraud Regressions

In this section, we use regression analysis to test our abnormal disclosure hypotheses using an unbalanced panel. As placebo tests, we consider not only disclosures in the year of an AAER, but also in the year prior and the year after the AAER. We expect a strong identifying signal only during the years of fraudulent activity, and not in the years prior to or after the alleged fraud periods. This approach allows us to fully understand the disclosure life cycle of fraudulent firms.

Table 5 displays the results of OLS regressions in which the dependent variable is the firm's disclosure strategy. As indicated in the first column, the dependent variable is either fraud similarity or the industry similarity score.[9] In Panel A to C, we report results for the

[9] Readers interested in fraud detection might prefer an alternative specification where the AAER dummy is the dependent variable for convenience and the fraud similarity score is an independent variable. Although such a specification produces similar results as it affirms the positive link between the fraud similarity score and AAER violations, we do not focus on this specification in our main tables due to potential endogeneity concerns. In particular, the sequencing of events in our framework is such that disclosure is created at the

entire sample, for larger firms, and for smaller firms, respectively. Firm size is identified using median assets in each year. These regressions are conservative in the sense that identification only is based on within-firm variation (they include controls for firm and year fixed effects). Standard errors are adjusted for clustering by firm. We also include several controls including the implied economic state of the firm (the average Tobins q and profitability of the ten firms in the given year having the most similar MD&A disclosure as the given firm based on cosine similarities).[10]

Panels D to F consider three robustness tests. Panel D considers the out of sample period (2002 and later). Panel E considers additional controls for restatements, litigation, mergers, and uncertainty. Panel F considers results based on industry fixed effects instead of firm fixed effects.

[Insert Table 5 Here]

Panel A of Table 5 shows that firms engaged in alleged fraud have significantly higher fraud profile similarities. This coefficient has a t-statistic of 6.58, and is significant well beyond the 1% level. The results for industry similarity are not significant (a t-statistic of 0.8). We note again that these regressions are based on stringent within-firm identification. The results for fraud similarity confirm the intuition established in the discussion of Figure 3, where we find that firms involved in fraud become more similar to other firms that committed fraud, but only in the years they are allegedly committing fraud. This suggests that these disclosures are likely related to commitment of the fraud itself.

Panels B and C of Table 5 show that fraud profile similarity is robust at the 1% level for both large and small firms. We also continue to find that industry similarity is not significant. We thus focus our attention on fraud profile similarity for the remainder of our study and conclude that fraudulent firms produce verbal disclosures that have a strong common component that cannot be explained by industry, size and age (ISA peers).

end of a fiscal year. Thus, the commission of fraud in the given fiscal year causes the disclosure to potentially have an abnormal component when the managers later summarize their fiscal year's performance, and not vice-a-versa. This indicates that the use of the fraud similarity score as the dependent variable is the appropriate model.

[10] The implied Tobins q and profitability of peers is particularly well-suited to control for economic conditions facing the firm in this setting as these are the conditions implied by the disclosure itself.

Panel D, shows that our results remain robust during the out of sample period from 2002 to 2008. This test is particularly stringent, as the sample is smaller, and the impact of firm fixed effects on remaining degrees of freedom is more extreme. Nevertheless, the fraud similarity variable remains significant at the 5% level with a t-statistic of 2.31.

In Panel E, we further challenge our specification by including four additional control variables: restatements, litigation, uncertainty and mergers.[11] Although we do not display the coefficients for these variables in Panel E to conserve space, we do report the full set of coefficients in Table A1 of the Online Appendix of this study. The inclusion of these particular variables in Panel E raises the bar for our tests as it examines whether our results are potentially due to narrower effects that have been documented in other studies. The results in Panel E show that our results are highly robust, as the t-statistic for fraud profile similarity is roughly equal in Panels A and E.

Panel F shows that our results are also robust to replacing firm fixed effects with less stringent SIC-2 industry fixed effects. Not surprisingly, the results are stronger. This indicates that although our results are primarily driven by within-firm variation, variation across industries also goes in the same direction, further supporting our key hypotheses.

Table 6 uses the same framework as Table 5, except that we consider the future AAER dummy (a dummy that is one if the firm will be involved in an AAER in the next fiscal year) as an explanatory variable instead of the actual AAER dummy. As a result, we are implicitly testing if fraud similarity is elevated in the year prior to the fraud period. This allows us to test hypotheses predicting that disclosure will strictly relate to the act of committing fraud, and not to passive long term firm characteristics. We thus expect that the results should be substantially weaker than those in Table 5.

[**Insert Table 6 Here**]

[11] The restatement words variable is logarithm of one plus the number of times the word "restatement" appears in the firm's MD&A section of the firm's 10-K text. The litigation dummy is the logarithm of one plus the number of times the word "litigation" appears in the firm's MD&A section of the firm's 10-K text. These two controls are intended to maximize their ability to explain our results given that we report later that these particular words are significantly related to post-AAER firms. We control for uncertainty using the standard deviation of monthly stock returns from the previous year, and we also include a dummy that is equal to one if the given firm-year observation does not have adequate CRSP data to compute this variable. The acquisition dummy is one if the firm was an acquirer in a merger, or in an acquisition of assets transaction from SDC Platinum, in the previous year.

Table 6 shows uniformly weak and statistically insignificant links between fraud profile similarity and the future AAER dummy. These results are thus much weaker than those in Table 5. This reinforces the graphical depiction of the average fraud score in Figure 3, which shows that fraud scores are close to zero prior to the fraud period. We conclude that our evidence in Table 5 is strongly linked to the years that firms are allegedly engaged in fraud and our results cannot be explained by passive long-term firm characteristics.

[**Insert Table 7 Here**]

Table 7 is similar to Table 6, except we replace the future AAER dummy with the past AAER dummy. Hence, the dummy identifies firms that have committed fraud in the past, but are no longer committing fraud. The results of Table 7 are similar to those of Table 6 in that fraud profile similarity is not positively related to AAERs. In some specifications, we in fact find a negative link. This finding suggests that, after they are caught, firms might adopt disclosures that distance themselves from prior bad behavior. One can think of this result as the "Repentant Manager" hypothesis. Overall, these results further show that our results are not related to passive firm characteristics, and are unique to firms allegedly committing fraud.

5 Content Analysis

The results in the previous section support the conclusion that fraudulent firms have a strong common component to their disclosure that is unique to the specific years in which they commit fraud. However, these tests do not provide specific support for our hypotheses. In this section, we consider content analysis using the 75 verbal factors based on Latent Dirichlet Allocation (LDA) from Ball, Hoberg and Maksimovic (2013).[12] Then we report the interpretable vocabulary topics from LDA that distinguish firms involved in AAERs from non-AAER firms. We again focus on a difference-based framework that includes firm fixed effects. We thus identify the specific verbal topics that appear while firms are

[12] The number of topics is the only material input the researcher needs to specify when running LDA. We use 75 topics following Ball, Hoberg and Maksimovic (2013), who document that 75 topics best summarize the value relevant information in MD&As.

allegedly involved in fraud, as compared to the same firms in the years they are not involved in fraud.

We conclude this section by examining placebo tests and we consider the years before and after a firm is involved in fraud, and we conduct parallel analysis for SEC comment letters. This latter test examines if our results for AAERs are related to the information in reviews conducted by the Division of Corporate Finance of the SEC, which comments on verbal disclosure such as MD&A.

5.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is based on the idea that a corpus of documents can be represented by a set of topics. LDA has been used extensively in computational linguistics and is replicable. It has the added benefit that it does not require researcher prejudgment. That is, the researcher cannot force assumptions in the model, as the derived topics are derived using automation.

The approach is commonly referred to as a "bag-of-words" technique because the relative frequency of words in a document is centrally important, but not their specific ordering. A particular topic can be characterized as a distribution over a common vocabulary of words where the relative probability weight assigned to each word indicates its relative importance to that topic. For example, the words Oil and Electricity might be important to topics associated with Natural Resources and Manufacturing, but one might expect Oil to dominate Electricity in the Natural Resources topic, while the opposite might be true for the Manufacturing topic.

We model each document as a mixture distribution over the topics rather than the words to better represent their content. Intuitively, the weight of each topic that is assigned to a particular document reflects its relative importance to the document. For example, the MD&A sections of British Petroleum and General Motors would both be expected to use the word "oil" but the documents might be expected to place greater emphasis respectively on the Manufacturing topics and Natural Resources topics respectively - both of which place high relative weight on this word. Hence the corresponding document generation process is assumed to arise from an underlying topic distribution, not an individual word distribution

LDA was developed by Blei, Ng and Jordan (2003) to provide an analytic framework that allows one to estimate the topic densities from a corpus of documents. We provide only a brief summary here, and refer readers to these articles for more detail. For our purposes, LDA is a generalization of factor analysis (used in numerical data) to textual data. It uses Gibbs Sampling and likelihood analysis and discovers clusters of text ("topics") that frequently appear in a corpus. We use the LDA topics from Ball, Hoberg and Maksimovic (2013), which were generated using the metaHeuristica software program.

LDA generates two detailed data structures. The first data structure is the set of word-frequency distributions for each topic. For LDA with 75 topics, this data structure contains 75 word lists with corresponding word frequencies. The word list also includes commongrams, which are 2-3 word phrases that also appear frequently in paragraphs that load highly on each topic. As do Ball, Hoberg and Maksimovic (2013), we fit the LDA model using only the first year of our sample (1997) to ensure there is no look ahead bias in the regressions that use LDA text.

The second data structure quantifies the extent to which each of the 75 topics is discussed in individual MD&As. These firm-year variables are commonly referred to as "topic loadings". For each firm in each year, LDA provides a vector of length 75 stating the extent to which the given firm's MD&A discusses each of the 75 topics.[13]

We use these two LDA-generated data structures to refine our understanding of disclosure during episodes of fraud. Our first objective is to provide a data structure for each topic that can be used to interpret each topic's content. We provide two computer generated resources. The first is the list of highest frequency commongrams, which provide a simple list of key phrases that associate with the topic. The second is a complete "representative paragraph", which is a paragraph that best represents the content that is typical among firms that use the topic.

We compute representative paragraphs by first extracting the 1000 paragraphs that have the highest cosine similarity with the probability-weighted word list associated with

[13] Generating the database of topic loadings is achieved by projecting the distribution of text for any given MD&A on the 75 vectors representing the distribution of text for each of the topics. See Ball, Hoberg and Maksimovic (2013) for details regarding this regression-based approach. This allows us to build a database of topic loadings for our entire sample 1997 to 2011 using the topic vocabularies from 1997 as discussed above.

each topic. The second step is to sort these paragraphs by their document lengths, and extract the middle tercile, which contains the set of typical length representative paragraphs. Among these 333 candidate paragraphs, we then define the displayed "representative paragraph" as the paragraph with the highest total similarity to the other 332 paragraphs in this set. This calculation is akin to computing network centrality, and hence the chosen "representative paragraph" should indeed be representative of the content associated with the given topic (thus making the topic more easily interpretable).

We then use the panel database containing the 75 numeric topic loadings for each firm in each year, and estimate regressions to infer which of the 75 verbal topics are most related to abnormal disclosures during periods of fraud, relative to disclosures the same firms make during years they are not allegedly committing fraud. The topics that are significantly different can then be interpreted and discussed regarding their potential consistency with our hypotheses.

5.2 Abnormal Content in AAER-Years

Table 8 displays the results of 75 regressions that treat each of the LDA topic loadings as a dependent variable. We only report the subsample of results for which the AAER dummy is significantly different from zero at the 5% level or better.[14] We control for year and firm fixed effects, and all standard errors are clustered by firm. Because we control for firm fixed effects, all reported links between fraud-years and LDA factors are conservative and based on within-firm identification. The first column reports the top 5 commongrams associated with each topic, and the second column displays the representative paragraph associated with each topic. These columns are designed to aid in the interpretation of each topic's content. We focus on the AAER dummy as the independent variable of interest in the final column.

[Insert Table 8 Here]

The table shows that twelve of the 75 topics are significantly linked to AAER years. These twelve topics are abnormally disclosed by firms involved in fraud relative to the same

[14] To be conservative, we do not report 10% level significant results given the number of specifications.

firms in non-AAER years (thus controlling for unobserved firm-specific variables). A negative coefficient indicates under-disclosure, and a positive coefficient indicates abnormally high levels of disclosure. We note that each topic is well-described by its representative paragraph and list of frequent commongrams, and we remind the reader that both are generated automatically as described above, and are thus not subjected to researcher prejudice. We also note that finding twelve significant topics, many significant at the 1% level, is well beyond what one would expect by chance for 75 topics.

5.2.1 Hypotheses Based on Fraud Verbal Disclosure Incentives

We next interpret the results in Table 8 through the lens of our three text-specific hypotheses (H1A, H1B, H1C) and the two fraud initiation hypotheses from the literature (H2A and H2B). We note that some hypotheses, especially H1A and H1B, are most directly examined by specifically considering expense or revenue fraud.

Regarding H1A, managerial incentives to conceal details of fraudulent accounting, row (1) is supportive. The representative paragraph shown in row (1) states for example:

> *"The increase was due primarily to the increase in sales, the decreased percentage of the general and administrative expenses and the decrease in depreciation and amortization expense. ... The decrease was due primarily to the increased costs of operating the company-owned restaurants."*

This paragraph explicitly explains the details underlying the firm's performance, and Table 8 shows that it is negatively related to the fraud dummy at better than the 1% level. This suggests that firms committing fraud disclose less information regarding details that explain how their performance arises, which supports H1A directly.

Regarding H1B, managerial incentives to grandstand growth and strong performance, we see direct support in Table 8. Firms appear to grandstand their growth as shown in row (11) on page 2 of the table, as they disclose significantly more of an LDA topic with a representative paragraph that touts the firm's growth. For example, this paragraph begins with the statement:

> *"The company's business has grown significantly since its inception..."*

We also note that hypothesis H1B is more directly tested in the subsamples of revenue fraud and expense fraud, which we consider in the next section. We note that the results in that section are also supportive of H1B.

Regarding H1C, managerial incentives to avoid references to themselves in the presence fraudulent accounting, we observe direct support in row (2). The representative paragraph for this topic notes that managers often discuss their plans for the future in MD&A and associate the plans with references to themselves. For example, the paragraph in row (2) starts with:

> *"Since joining the company in January 1998, the new chief executive officer, along with the rest of the company's management team has been developing a broad operational and financial restructuring plan..."*

Table 8 shows that this kind of self-reference is less likely to occur when the firm is committing fraud, which directly supports H1C as managers prefer to disassociate with the firm's disclosure when they are committing fraud, likely to insulate themselves from the fallout should the fraud be discovered in the future.

5.2.2 Hypotheses Based on Fraud Initiation Incentives

We next consider H2A and H2B. Row (4) of Table 8 supports H2A, and shows that firms under-disclose discussions of liquidity when they are committing fraud. We also note that these regressions control for firm and year fixed effects. Hence, these results are not attributable to poorly performing firms in general and they also cannot be explained by long-term tendencies of some firms to discuss financial market liquidity. Regarding H2B, the hypothesis that fraud is related to incentives surrounding mergers and acquisitions, we find that row (5) provides direct support.

5.2.3 Placebo Tests Based on Non-Fraud Periods

We next consider placebo tests and examine whether the above results are robust to replacing the AAER dummy with a pre-AAER dummy or a post-AAER dummy. If our results look materially the same in these placebo periods, that would indicate that our results would not be uniquely attributable to the periods during which firms actually commit fraud. We report the results in Table 9. In particular, we report all topics that are significantly related to actual fraud years in the first column, pre-AAER years in the second column, and post-AAER years in the third column. The first column thus reports exactly the same results as Table 8 for comparison, and we draw attention to the second and third columns.

[**Insert Table 9 Here**]

Table 9 shows that all of our central results only exist in the actual AAER sample, and not in the pre-AAER or the post-AAER sample. Regarding the post-AAER sample, not a single coefficient is significant with the same sign. Regarding the pre-AAER sample, only the result for legal proceedings has the same sign and is significant. Given the number of topics that are related to AAERs in Table 8, these results suggest that our findings are indeed unique to the years during which firms are actually committing fraud, as required by our key hypotheses.

5.3 Revenue and Expense Fraud

We next reexamine the tests in Table 8 specifically for revenue fraud and expense fraud. These tests specifically allow us to more precisely examine hypotheses H1A (incentives to conceal details) and H1B (incentives to grandstand). The separate consideration of revenue and expense fraud for these hypotheses is relevant as each hypothesis further predicts that different types of disclosure will be especially high or low when the manager is committing revenue fraud or expense fraud. For example, given that a manager is committing revenue fraud, H1B predicts that managers will specifically grandstand the strong revenue growth. In contrast, when the manager is committing expense fraud, H1B predicts more grandstanding of the manager's cost management.

[**Insert Table 10 Here**]

Table 10 reruns the same specifications in Table 8, which is based on all frauds, specifically for revenue and expense frauds. Regarding hypothesis H1B (grandstanding), row (3) provides direct evidence. The corresponding representative paragraph begins with:

> *"Revenues increased by $29.9 million, or approximately 27.4%, to $139.1 million in 1997 from $109.2 million in 1996."*

Managers committing revenue fraud abnormally disclose longer discussions touting their revenue performance. Consistent with H1A, their MD&As also under-disclose details regarding how their performance arises. This latter conclusion is supported by both row (1) and row (4), which directly indicate the disclosure of fewer details explaining the performance of the firm. For example, the representative paragraph associated with row (1) starts with:

> *"Income from operations for 1997 totaled $974,549, an increase of $121,707 (14.3%) from 1996. The increase was due primarily to the increase in sales, the decreased percentage of the general and administrative expenses and the decrease in depreciation and amortization expense."*

The results for expense fraud also support both H1A and H1B. Regarding H1B, row (7) shows that managers committing expense fraud disclose abnormally high levels of disclosure relating to R&D expense performance. The representative paragraph starts with:

> *"Research and development expenses increased 20.7% to $6,006,000 in 1996, and increased as a percentage of net sales to 10.0% in 1996 from 6.1% in 1995. The increases in research and development expenses were primarily due to the expansion of the research and development staff, and expenses associated with its research and development facility."*

This is consistent with grandstanding to convince investors that the firm has strong growth options consistent with having a vibrant R&D program. These firms also disclose fewer details explaining their performance, as illustrated by the negative coefficient in row (1) regarding broad performance details and row (10) regarding specific cost reduction details.

5.4 Fog Index

In this section, we test whether managers use language that is difficult to read in order to obfuscate their disclosures. We compute the Gunning Fog Index for each firm's MD&A in each year, and consider regressions analogous to those in Table 5 where the Gunning Fog Index is the dependent variable. Under this hypothesis, we expect the AAER dummy to be a positive and significant predictor of the Gunning Fog Index. The formula for the

Gunning Fog Index is $0.4[\frac{\#words}{\#sentences} + \frac{\#complexwords}{\#words}]$, where complex words are those with three or more syllables. We also consider the Automated Readability Index and the Flesch Kinkaid Index for robustness.

[Insert Table 11 Here]

The results are reported in Table 11. Panel A, which is based on the AAER year, does not support the hypothesis that managers use complex text when they are involved in AAERs. For two of the three indices, the AAER dummy is negative and significant. The coefficient becomes positive and significant only in Panel C, which is based on the post-AAER year. The likely explanation is that once the AAER becomes public, firms disclose the legal implications of the AAER itself, and the use of legal jargon increases the difficulty of reading the document and hence the various fog indices.

Although we do not find evidence of obfuscation, we note that the obfuscation hypothesis is broader than MD&A and it is possible that obfuscation occurs in other disclosures.

5.5 Robustness

As an additional robustness examination, we identify the individual words that are used more aggressively by AAER firms. These words are identified based on word-by-word tests of differences in each word's relative usage among AAER firms versus non-AAER firms. The details of this analysis are not reported here but are available in Table A2 of our online appendix. Table A2 shows that AAER years are often linked to restatements, which indicates a history of poor accounting beyond the AAER itself. We also observe that AAER firms disclose more information about acquisitions and international vocabulary including region and country names such as Africa and Brazil. It is possible that more difficult to trace international transactions might facilitate fraudulent accounting. Firms involved in AAERs also disclose more vocabulary indicative of uncertainty and speculation: "believe", "feasibility", "fluctuating", and "instability".

Our general conclusion, however, is that individual words are more difficult to interpret than are the results for LDA discussed previously. This comparison thus highlights how word-clustering methods like LDA can add clarity to content analysis. We also report single word results for pre-AAER and post-AAER firms in the Online Appendix tables A3 and A4. These tables also confirm that AAER years are unique. Table A4 confirms that firms involved in AAERs disclose information about the AAER itself after the AAER investigation is made public. We also present a list of the top 25 most representative AAERs in Table A5, which lists the AAERs that have the highest fraud similarity scores.

In online appendix A6, we examine which topics are disclosed abnormally high or low when SEC Comment Letters are issued. Because the comment letter process is the process of review by the SEC, this test thus examines if our aforementioned results are driven by known SEC evaluation processes, or if they are instead linked to incentives as suggested in the hypotheses section of the paper. We thus consider the same regression model as in Table 8, but we replace the AAER dummy with a dummy indicating whether the firm received a comment letter from the SEC relating to its MD&A disclosure in the given year. The table shows that there is little overlap between the disclosure of firms receiving comment letters and those committing fraud. The results thus support the conclusion that our results are likely driven by incentives to commit fraud, and are not artifacts of the SEC review process itself.

6 Equity Market Liquidity

This section more deeply considers hypothesis H2A: the proposed link between fraudulent firm disclosure, equity market liquidity and equity issuance. We examine the more specific hypothesis that managers might commit fraud to get access to an artificially lower cost of capital (see Dechow, Sloan and Sweeney (1996), Povel, Singh and Winton (2007) and Wang, Winton and Yu (2010)). We consider whether, following exogenous negative shocks to equity market liquidity, managers are more likely to commit fraud and produce disclosure with a higher fraud similarity score, likely to inflate their odds of issuing equity.

We consider the Coval and Stafford (2007) and Edmans, Goldstein, Jiang (2012) forced mutual fund selling shock as an exogenous negative shock to equity market liquidity. As this measure of forced mutual fund selling is not sector-specific, and only affects equities, it is a direct shock to equity market liquidity. The authors also find that the effects of this shock can be long lasting, as much as two years. We examine regressions in which the dependent variable is the fraud profile similarity score or the AAER dummy, and the mutual fund selling shock is a key independent variable. If improving the odds of issuing equity is a strong motive for fraud that drives the common verbal disclosures made by fraudulent firms, the prediction is that negative shocks to equity market liquidity should result in increases in the fraud profile similarity score and the AAER dummy. This prediction arises from the assumption that the incentive to commit fraud increases when liquidity conditions deteriorate.

[**Insert Table 12 Here**]

The results are presented in Table 12 Panel A and we include firm and year fixed effects. The results support our prediction that negative shocks to equity market liquidity lead firms to produce disclosure with higher fraud profile similarity scores. Moreover, the same firms are more likely to be involved in an AAER in these years. These results are highly significant.

In panel B, we examine regressions in which the dependent variable is equity issuance, and the key independent variable is the fraud profile similarity. We again include firm and year fixed effects. As indicated in the first column, we consider equity issuance measured two ways: Compustat equity issuance/assets and SDC Platinum public SEO proceeds/assets. Our hypothesis is that if fraudulent disclosure is made to inflate the odds of issuing equity, and if the market is not fully aware of this link, then increased fraud profile similarity should predict more equity issuance.

We note, however, that these panel B regressions are only suggestive, as the link between disclosure and equity issuance is potentially endogenous. We are not aware of any instruments for increased fraud profile similarity disclosure that are unrelated to liquidity. The results are consistent with the conclusion that firms with high fraud profile similarity issue more equity than firms with lower scores. Overall, our results in Panel A suggest a potential causal link between poor equity market liquidity and elevated levels of fraud profile similarity. Panel B is consistent with a non-causal link to equity issuance.

7 Conclusions

We first examine if firms committing fraud produce abnormal disclosure that is common among firms committing fraud. We define abnormal disclosure as that which cannot be explained by industry peers of similar size and age, or as disclosure that cannot be explained by firm fixed effects and various controls. We find such an abnormal component among fraudulent firms, and a any firm's verbal similarity to this abnormal vocabulary predicts ex-post fraud both in sample and out of sample. The results are economically large. Firms in the lowest fraud vocabulary similarity decile commit fraud at a rate of 0.5%, while those in the highest decile commit fraud at a rate of 3.7%. These results are also robust to controlling for firm fixed effects, and we continue to find strong results even when compared to the same firms before and after the AAER. These results suggest that disclosures are revised as a firm evolves from a pre-fraud firm, to a firm involved in fraud, and later to a firm that has been revealed as having committing fraud.

Having established the presence of abnormal disclosure, we turn our attention to understanding why. Our tests reveal a link between fraudulent firms and the under-reporting of details explaining how the firm's accounting performance arises, and grandstanding the firm's growth potential and its strong performance. These results suggest that managers might respond to incentives to conceal details that might increase detection, and incentives to grandstand growth and performance to increase the positive impact the manipulation has on the firm's outcomes. Finally, we also find evidence that fraudulent managers disclose in such a way to disassociate their own names and reputations from the fraudulent performance. These three hypotheses based on verbal content (concealing details, grandstanding performance, and self disassociation) are novel, especially as they relate to MD&A disclosures in the 10-K.

We also find specific textual support for two hypotheses noted in the existing literature: managers commit fraud to improve their odds of raising capital, and managers commit fraud to improve acquisition terms.

Finally, we consider negative exogenous shocks to equity market liquidity in order to further challenge the hypothesis that fraud is driven by incentives to increase the odds of raising capital. We find that these exogenous liquidity shocks are associated with increased use of the abnormal vocabulary that is common among fraudulent firms, and also with increased rates of ex-post fraud. These results are consistent with a causal role for equity market liquidity in determining the amount of abnormal disclosure that firms produce. Our results have implications for improving the ability to detect fraud, and to further understand the motives that are most salient in driving managers to commit fraud and produce abnormal disclosure.

References

Antweiler, Werner, and Murray Frank, 2004, Is all that talk just noise? The information content of internet stock message boards, *Journal of Finance* 52, 1259–1294.

Ball, Christopher, and Gerard Hoberg, and Vojislav Maksimovic, 2013, Disclosure Informativeness and the Tradeoff Hypothesis: A Text-Based Analysis, University of Maryland Working Paper.

Beck, Thorsten, Asli Demirguc-Kunt, Asli and Vojislav Maksimovic, Vojislav,2005, Financial and Legal Constraints to Growth: Does Firm Size Matter? *Journal of Finance* 60, 137–77.

Beneish, Messod, 1997, Detecting GAAP Violation: Implications for Assessing Earnings Management among Firms with Extreme Financial Performance, *Journal of Accounting and Public Policy* 16, 271–309.

Beneish, Messod, 1999, The Detection of Earnings Manipulation, *Financial Analysts Journal* 55, 24–36.

Beneish, Messod, 1999, Incentives and Penalties Related to Earnings Overstatements that Violate GAAP, *The Accounting Review* 74, 425–457.

Blei, D. M., Ng, A. Y. and Jordan, M. I., 2003, Latent Dirichlet allocation, *Journal of Machine Learning Research* 3, 993–1002.

Brown, Stephen V., and Jennifer Wu Tucker, 2011, Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications, *Journal of Accounting Research* 49, 309–346.

Bryan, S. H., 1997, Incremental Information Content of Required Disclosures Contained in Management Discussion and Analysis, *The Accounting Review* 72, 285–301.

Burns, Natasha, and Simi Kedia, 2006, The impact of performance-based compensation on misreporting, *Journal of financial economics* 79, 35–67.

Cimiano, Phillip, 2010, *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*, Springer, New York.

Cole, C. J., and C. L. Jones, 2005, Management Discussion and Analysis: A Review and Implications for Future Research, *Journal of Accounting Literature* 24, 135–74.

Coval, Joshua, and Erik Stafford, 2007, Asset fire sales (and purchases) in equity markets, *Journal of Financial Economics* 86, 479-512.

Darrough, Masako N., 1993, Disclosure policy and competition: Cournot vs. Bertrand, *Accounting Review* 534–561.

Dechow, Patricia, and Weili Ge, and Chad Larson, and Richard Sloan, 2011, Predicting Material Accounting Misstatements, *Contemporary Accounting Research* 28, 17–82.

Dechow, Patricia, and Weili Ge, and Catherine Schrand, 2010, Understanding earnings quality: A review of the proxies, their determinants and their consequences, *Journal of Accounting and Economics* 2, 344–401.

Dechow, Patricia, and Richard Sloan, and Amy Sweeney, 1996, Causes and Consequences of Earnings Manipulation: An Analysis of Firms Subject to Enforcement Actions by the SEC, *Contemporary Accounting Research* 13, 1–36.

Devenow, Andrea, and Ivo Welch, 1996, Rational Herding in Financial Economics, *European Economic Review* 40, 603–615.

Dye, Ronald A., and Sri S. Sridhar, 1995, Industry-wide disclosure dynamics, *Journal of accounting research* 157–174.

Dyck, Alexander, and Adair Morse, and Luigi Zingales, 2010, Who Blows the Whistle on Corporate Fraud?, *Journal of Finance* 65, 2213–53.

Edmans, A., I. Goldstein, and W. Jiang, 2012, The Real Effects of Financial Markets: The Impact of Prices on Takeovers, *The Journal of Finance* 67, 933–971.

Erickson, M., and S.W. Wang, 1999, Earnings management by acquiring firms in stock for stock mergers, *Journal of Accounting and Economics* 27, 149–176.

Feldman, R., S. Govindaraj, J. Livnat, and B. Segal, 2010, Management's Tone Change, Post Earnings Announcement Drift and Accruals, *Review of Accounting Studies* 15, 915–53.

Feroz, Ehsan, and Kyungjoo Park, and Vector Pastena, 1991, The Financial and Market Effects of the SEC's Accounting and Auditing Enforcement Releases, *Journal of Accounting Research* 29, 107–142.

Goldman, Eitan, and Steve Slezak, 2006, An equilibrium model of incentive contracts in the presence of information manipulation, *Journal of Financial Economics* 80, 603–26.

Hanley, Kathleen, and Gerard Hoberg, 2010, The information content of IPO prospectuses, *Review of Financial Studies* 23, 2821–2864.

Hanley, Kathleen, and Gerard Hoberg, 2012, Litigation risk and the underpricing of initial public offerings, *Journal of Financial Economics* 103, 235–254.

Healy, Paul, and James Wahlen, 1999, A Review of the Earnings Management Literature and its Implications for Standard Setting, *Accounting Horizons* 13, 365–383.

Hoberg, Gerard, and Vojislav Maksimovic, 2012, Redefining Financial Constraints: a Text-Based Analysis, *Review of Financial Studies*, Forthcoming.

Hoberg, Gerard, and Gordon Phillips, 2010, Product market synergies in mergers and acquisitions: A text based analysis, *Review of Financial Studies* 23, 3773–3811.

Hoberg, Gerard, and Gordon Phillips, 2012, New dynamic product based industry classifications and endogenous product differentiation, *University of Maryland Working Paper.*

Hughes, Patricia J., and Anjan V. Thakor, 1992, Litigation risk, intermediation, and the underpricing of initial public offerings, *Review of Financial Studies* 5, 709–742.

Johnson, Shane, and Harley Ryan and Yisong Tian, 2009, Managerial Incentives and Corporate Fraud: The Sources of Incentives Matter, *Review of Finance* 1, 115–145.

Karpoff, Jonathan, and Scott Lee and Gerald Martin, 2008, The Consequences to Managers for Financial Misrepresentation, *Journal of Financial Economics* 88, 193–215.

Karpoff, Jonathan, and Scott Lee and Gerald Martin, 2008, The Cost to Firms of Cooking the Books, *Journal of Quantitative and Financial Analysis* 43, 581–612.

Kedia, Simi, and Thomas Philippon, 2009, The Economics of Fraudulent Accounting, *Review of Financial Studies* 22, 2169–2199.

Kothari, S. P., Xu Li, and James E. Short, 2009, The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study Using Content Analysis, *The Accounting Review* 84, 1639–70.

Li, Feng, 2010, Information Content of the Forward-Looking Statements in Corporate Filings: A Naive Bayesian Machine Learning Approach, *Journal of Accounting Research* 48, 1049–1102.

Li, Feng, 2008, Annual Report Readability, Current Earnings, and Earnings Persistence, *Journal of Accounting and Economics* 45, 221–47.

Loughran, Tim, and Bill McDonald, 2011, When is a liability not a liability? Textual analysis, dictionaries, and 10-ks, *Journal of Finance* 66, 35–65.

Povel, Paul, and Rajdeep Singh, and Andrew Winton, 2007, Booms, Busts, and Fraud, *Review of Financial Studies* 20, 1219–1254.

Sebastiani, Fabrizio, 2002, Machine learning in automated text categorization, *acmcs.*

Skinner, Douglas, and Richard Sloan, 2002, Earnings Surprises, Growth Expectations, and Stock Returns or Don't Let an Earnings Torpedo Sink Your Portfolio, *Review of Accounting Studies* 7, 289–312.

Tetlock, Paul C., 2007, Giving content to investor sentiment: The role of media in the stock market, *Journal of Finance* 62, 1139–1168.

Wang, Tracy, 2013, Corporate Securities Fraud: Insights from a New Empirical Framework, *Journal of Law and Economics* 29, 535–568.

Wang, Tracy, Andrew Winton, Xiaoyun Yu, 2010, Corporate Fraud and Business Conditions: Evidence from IPOs, *Journal of Finance* 65, 2255–2292.

Table 1: Summary Statistics

Summary statistics are reported for our sample of 49,039 observations based on annual firm observations from 1997 to 2008. The AAER dummy is one if an AAER action indicates that the firm was involved in fraudulent activity in the current year. The industry similarity score is the raw cosine similarity of the given firm's MD&A disclosure and that of its industry-size-age peers. These peers are identified by sorting firms in each two digit SIC code first into above and below median firm sizes, and then into above and below median firm ages for each group. Median size and age are computed separately for each year. A higher figure indicates that the given firm has disclosure that is highly similar to its industry peers. To compute the fraud similarity score, we first compute each firm's abnormal disclosure as its raw disclosure minus the average disclosure of its industry-size-age peers. The fraud similarity score is then the cosine similarity of the given firm's abnormal disclosure and the average abnormal disclosure of all firms involved in AAERs in the sample period 1997 to 2001. We use these earlier years of our sample to identify the vocabulary of firms allegedly committing fraud so that we can consider out of sample analysis for the later years in our sample 2002 to 2008. Log Sales is the natural logarithm of Compustat sales. Operating Income/Sales is Compustat operating income before depreciation scaled by sales. R&D/sales and CAPX/sales are Compustat values of R&D and capital expenditures scaled by sales. All ratios are winsorized at the 1% and 99% level, and any values of operating income/sales less than minus one are set to minus one.

Variable	Mean	Std. Dev.	Minimum	Median	Maximum
AAER Dummy	0.015	0.120	0.000	0.000	1.000
Industry Similarity Score	0.667	0.080	0.410	0.671	0.839
Fraud Similarity Score	0.002	0.077	-0.191	-0.002	0.251
Log Sales	4.917	2.127	0.001	4.866	12.326
Operating Income/Sales	-0.006	0.353	-1.000	0.081	0.703
R&D/Sales	0.190	0.770	0.000	0.000	11.230
CAPX/Sales	0.123	0.345	0.000	0.037	9.276

Table 2: Pearson Correlation Coefficients

Pearson Correlation Coefficients are reported for our sample of 49,039 observations based on annual firm observations from 1997 to 2008. See Table 1 for the description of our key variables.

Row	Variable	AAER Dummy	Fraud Similarity Score	Industry Similarity Score	Log Sales	Operating Income/ Sales	R&D Sales
				Correlation Coefficients			
(1)	Fraud Similarity Score	0.082					
(2)	Industry Similarity Score	0.026	0.090				
(3)	Log Sales	0.070	-0.005	0.061			
(4)	Operating Income/Sales	0.022	-0.026	-0.040	0.522		
(5)	R&D/Sales	-0.012	0.044	0.081	-0.302	-0.518	
(6)	CAPX/Sales	-0.010	-0.002	0.042	-0.145	-0.156	0.195

Table 3: AAER Timeseries Statistics

The table reports time series statistics for our sample of 49,039 observations based on annual firm observations from 1997 to 2008. The AAER dummy is one if an AAER action indicates that the firm was involved in fraudulent activity in the current year.

Row	Year	Number AAER Firm Years	Number of Firms in Sample	Fraction AAER Firm Years
1	1997	28	4670	0.006
2	1998	48	4663	0.010
3	1999	80	4727	0.017
4	2000	110	4647	0.024
5	2001	125	4406	0.028
6	2002	104	4173	0.025
7	2003	80	4009	0.020
8	2004	68	3915	0.017
9	2005	46	3522	0.013
10	2006	17	3396	0.005
11	2007	10	3420	0.003
12	2008	4	3491	0.001

Table 4: AAERs versus Fraud Similarities and Industry Similarity Deciles

The table displays decile statistics for our sample of 49,039 observations based on annual firm observations from 1997 to 2008. Within each year, firms are sorted into deciles based on their fraud similarities (first two columns) and based on their industry similarity scores (latter two columns). The fraction of firms involved in AAERs is then reported for each decile group. See Table 1 for the description of our key variables.

Decile	Fraud Similarity Score	Fraction AAER Firm Years	Industry Similarity Score	Fraction AAER Firm Years
		Panel A: Full Sample (1997-2008)		
1	-0.124	0.005	0.514	0.010
2	-0.076	0.007	0.585	0.012
3	-0.050	0.008	0.617	0.012
4	-0.030	0.011	0.641	0.012
5	-0.011	0.010	0.662	0.012
6	0.007	0.011	0.682	0.015
7	0.027	0.014	0.702	0.017
8	0.050	0.020	0.724	0.013
9	0.081	0.023	0.750	0.017
10	0.147	0.037	0.792	0.027
		Panel B: Out of Sample (2002-2008)		
0	-0.112	0.007	0.519	0.009
1	-0.069	0.008	0.587	0.008
2	-0.045	0.009	0.616	0.008
3	-0.026	0.014	0.639	0.009
4	-0.010	0.012	0.657	0.010
5	0.007	0.010	0.676	0.013
6	0.024	0.008	0.696	0.016
7	0.044	0.018	0.718	0.011
8	0.070	0.017	0.745	0.018
9	0.129	0.022	0.789	0.024

Table 5: Disclosure Outcome Regressions (AAER-year)

In Panels A to C, the table reports our baseline OLS regressions for our sample of 49,039 firm-year observations based on annual firm observations from 1997 to 2008. These baseline regressions are estimated with year and firm fixed effects. The dependent variable is based on a firm-year's disclosure in its 10-K and varies by row as indicated. For all Panels, standard errors are clustered by firm and t-statistics are in parentheses. See Table 1 for the description of our key variables. The AAER dummy is our primary variable of interest and is one if an AAER action indicates that the firm was involved in fraudulent activity in the current year. To control for the economic conditions associated with the information in a given firm's MD&A, we additionally include controls for the average Tobins Q and Operating Income/Sales for the ten firms with MD&A Sections that are most similar to the given firm (these ten firms are those with the highest cosine similarity between their MD&A and that of the given firm in the given year). Panels D to F consider various robustness tests regarding the baseline model in Panel A. Panel D repeats the test in Panel A but only for our out of sample period including 25,926 annual firm observations from 2002 to 2008. These tests are out of sample because the base vocabulary used to compute fraud similarity is fitted only using the earlier subsample from 1997 to 2001. One observation is one firm in one year. Panel E repeats the test in Panel A but adds three additional control variables aimed at challenging whether our results are due to narrower effects that have been documented in other studies but that might be better measured using text: a restatement variable, a litigation variable, a control for uncertainty, a control for inadequate CRSP data to compute uncertainty, and an acquisition dummy. Although we include these variables in the model, we do not report the additional coefficients to conserve space (see the Online Appendix for the presentation of those coefficients). Panel F repeats the test in Panel A but replaces the firm fixed effects with SIC-2 industry fixed effects.

Row	Dependent Variable	AAER Dummy	Operating Income /Sales	R&D /Sales	CAPX /Sales	Log Sales	MD&A Peer Implied Tobins Q	MD&A Peer Implied OI/sales	Obs.
				Panel A: Entire Sample					
(1)	Fraud Profile Sim.	0.029	0.000	0.002	0.004	0.005	0.003	–0.001	49,039
		(6.58)	(–0.01)	(1.40)	(2.68)	(5.26)	(11.63)	(–3.66)	
(2)	Industry Similarity	0.003	–0.012	0.003	0.004	0.009	–0.001	0.000	49,039
		(0.77)	(–4.87)	(3.86)	(2.32)	(9.41)	(–2.65)	(2.88)	
				Panel B: Above Median Firm Size Only					
(3)	Fraud Profile Sim.	0.026	0.020	0.046	0.012	0.008	0.004	–0.001	24,523
		(4.93)	(3.57)	(2.49)	(3.58)	(4.76)	(8.37)	(–2.14)	
(4)	Industry Similarity	–0.001	–0.003	–0.005	–0.002	0.004	–0.001	0.000	24,523
		(–0.13)	(–0.41)	(–0.32)	(–0.52)	(2.26)	(–2.88)	(1.35)	
				Panel C: Below Median Firm Size Only					
(5)	Fraud Profile Sim.	0.031	–0.004	0.001	0.003	0.005	0.003	–0.001	24,516
		(3.44)	(–1.46)	(1.06)	(1.84)	(3.95)	(7.92)	(–3.02)	
(6)	Industry Similarity	0.008	–0.017	0.004	0.005	0.011	0.000	0.000	24,516
		(1.09)	(–6.13)	(3.99)	(2.85)	(8.77)	(–1.58)	(2.46)	
				Panel D: Same as Panel A, but Out of Sample Years Only					
(7)	Fraud Profile Sim.	0.014	–0.006	0.002	0.004	0.003	0.003	–0.001	25,926
		(2.31)	(–1.66)	(1.31)	(1.95)	(2.10)	(4.03)	(–3.22)	
(8)	Industry Similarity	0.008	–0.014	0.003	0.008	0.009	–0.002	0.000	25,926
		(1.41)	(–3.56)	(2.47)	(2.72)	(5.80)	(–2.49)	(1.40)	
				Panel E: Same as Panel A, but Add Additional Controls					
(9)	Fraud Profile Sim.	0.029	0.002	0.001	0.004	0.004	0.003	–0.001	49,039
		(6.76)	(0.65)	(1.33)	(2.75)	(4.36)	(11.73)	(–3.49)	
(10)	Industry Similarity	0.003	–0.010	0.003	0.004	0.008	–0.001	0.001	49,039
		(0.78)	(–4.21)	(3.68)	(2.37)	(8.63)	(–2.71)	(3.06)	
				Panel F: Same as Panel A, but Replace Firm Effects with Industry Effects					
(11)	Fraud Profile Sim.	0.049	0.003	0.004	-0.003	0.001	0.006	-0.000	49,745
		(9.00)	(1.21)	(5.08)	(-2.25)	(2.21)	(14.33)	(-1.23)	0.031
(12)	Industry Similarity	0.005	-0.013	0.009	0.005	0.008	0.001	0.000	49,745
		(1.04)	(-6.27)	(11.23)	(2.64)	(19.45)	(2.45)	(2.65)	0.198

Table 6: Disclosure Outcome Regressions (Pre-AAER Disclosures)

In Panels A to C, the table reports our baseline OLS regressions for our sample of 49,039 firm-year observations based on annual firm observations from 1997 to 2008. These baseline regressions are estimated with year and firm fixed effects. The dependent variable is based on a firm-year's disclosure in its 10-K and varies by row as indicated. For all Panels, standard errors are clustered by firm and t-statistics are in parentheses. See Table 1 for the description of our key variables. The Future AAER dummy is our primary variable of interest and is one if the firm was involved in fraudulent activity in the year after the current year of the observation. To control for the economic conditions associated with the information in a given firm's MD&A, we additionally include controls for the average Tobins Q and Operating Income/Sales for the ten firms with MD&A Sections that are most similar to the given firm (these ten firms are those with the highest cosine similarity between their MD&A and that of the given firm in the given year). Panels D to F consider various robustness tests regarding the baseline model in Panel A. Panel D repeats the test in Panel A but only for our out of sample period including 25,926 annual firm observations from 2002 to 2008. These tests are out of sample because the base vocabulary used to compute fraud similarity is fitted only using the earlier subsample from 1997 to 2001. One observation is one firm in one year. Panel E repeats the test in Panel A but adds five additional control variables aimed at challenging whether our results are due to narrower effects that have been documented in other studies but that might be better measured using text: a restatement variable, a litigation variable, a control for uncertainty, a control for inadequate CRSP data to compute uncertainty, and an acquisition dummy. Although we include these variables in the model, we do not report the additional coefficients to conserve space (see the Online Appendix for the presentation of those coefficients). Panel F repeats the test in Panel A but replaces the firm fixed effects with SIC-2 industry fixed effects.

Row	Dependent Variable	AAER Dummy	Operating Income /Sales	R&D /Sales	CAPX /Sales	Log Sales	MD&A Peer Implied Tobins Q	MD&A Peer Implied OI/sales	Obs.
	Panel A: Entire Sample								
(1)	Fraud Profile Sim.	–0.001	0.000	0.002	0.004	0.005	0.003	–0.001	49,039
		(–0.18)	(–0.07)	(1.43)	(2.77)	(5.53)	(11.68)	(–3.73)	
(2)	Industry Similarity	–0.004	–0.012	0.003	0.004	0.009	–0.001	0.000	49,039
		(–0.71)	(–4.87)	(3.87)	(2.33)	(9.45)	(–2.65)	(2.88)	
	Panel B: Above Median Firm Size Only								
(3)	Fraud Profile Sim.	–0.001	0.020	0.047	0.013	0.009	0.004	–0.001	24,523
		(–0.12)	(3.53)	(2.51)	(3.61)	(4.96)	(8.66)	(–2.35)	
(4)	Industry Similarity	–0.010	–0.003	–0.005	–0.002	0.004	–0.001	0.000	24,523
		(–1.41)	(–0.42)	(–0.33)	(–0.52)	(2.25)	(–2.89)	(1.38)	
	Panel C: Below Median Firm Size Only								
(5)	Fraud Profile Sim.	–0.005	–0.004	0.001	0.003	0.005	0.003	–0.001	24,516
		(–0.39)	(–1.52)	(1.09)	(1.90)	(4.10)	(7.90)	(–3.02)	
(6)	Industry Similarity	0.001	–0.017	0.004	0.005	0.011	0.000	0.000	24,516
		(0.12)	(–6.15)	(4.00)	(2.87)	(8.83)	(–1.59)	(2.46)	
	Panel D: Entire Sample (Out of Sample Years Only)								
(7)	Fraud Profile Sim.	0.001	–0.006	0.002	0.004	0.003	0.003	–0.001	25,926
		(0.05)	(–1.69)	(1.32)	(2.00)	(2.19)	(4.06)	(–3.23)	
(8)	Industry Similarity	–0.016	–0.014	0.003	0.008	0.010	–0.002	0.000	25,926
		(–0.78)	(–3.57)	(2.47)	(2.74)	(5.82)	(–2.47)	(1.40)	
	Panel E: Same as Panel A, but Add Additional Controls								
(9)	Fraud Profile Sim.	0.001	0.001	0.001	0.004	0.004	0.003	–0.001	49,039
		(0.22)	(0.59)	(1.35)	(2.84)	(4.61)	(11.79)	(–3.56)	
(10)	Industry Similarity	–0.002	–0.010	0.003	0.004	0.008	–0.001	0.001	49,039
		(–0.34)	(–4.21)	(3.69)	(2.38)	(8.67)	(–2.70)	(3.06)	
	Panel F: Same as Panel A, but Replace Firm Effects with Industry Effects								
(11)	Fraud Profile Sim.	0.010	0.002	0.004	-0.003	0.001	0.006	-0.000	49,745
		(1.24)	(1.15)	(5.07)	(-2.14)	(2.98)	(14.38)	(-1.36)	0.025
(12)	Industry Similarity	-0.001	-0.013	0.009	0.005	0.008	0.001	0.000	49,745
		(-0.15)	(-6.28)	(11.23)	(2.64)	(19.54)	(2.48)	(2.64)	0.198

Table 7: Disclosure Outcome Regressions (Post-AAER Disclosures)

In Panels A to C, the table reports our baseline OLS regressions for our sample of 49,039 firm-year observations based on annual firm observations from 1997 to 2008. These baseline regressions are estimated with year and firm fixed effects. The dependent variable is based on a firm-year's disclosure in its 10-K and varies by row as indicated. For all Panels, standard errors are clustered by firm and t-statistics are in parentheses. See Table 1 for the description of our key variables. The Past AAER dummy is our primary variable of interest and is one if the firm was involved in fraudulent activity in the year prior to the current year of the observation. To control for the economic conditions associated with the information in a given firm's MD&A, we additionally include controls for the average Tobins Q and Operating Income/Sales for the ten firms with MD&A Sections that are most similar to the given firm (these ten firms are those with the highest cosine similarity between their MD&A and that of the given firm in the given year). Panels D to F consider various robustness tests regarding the baseline model in Panel A. Panel D repeats the test in Panel A but only for our out of sample period including 25,926 annual firm observations from 2002 to 2008. These tests are out of sample because the base vocabulary used to compute fraud similarity is fitted only using the earlier subsample from 1997 to 2001. One observation is one firm in one year. Panel E repeats the test in Panel A but adds three additional control variables aimed at challenging whether our results are due to narrower effects that have been documented in other studies but that might be better measured using text: a restatement variable, a litigation variable, a control for uncertainty, a control for inadequate CRSP data to compute uncertainty, and an acquisition dummy. Although we include these variables in the model, we do not report the additional coefficients to conserve space (see the Online Appendix for the presentation of those coefficients). Panel F repeats the test in Panel A but replaces the firm fixed effects with SIC-2 industry fixed effects.

							MD&A	MD&A	
			Operating				Peer	Peer	
	Dependent	AAER	Income	R&D	CAPX	Log	Implied	Implied	
Row	Variable	Dummy	/Sales	/Sales	/Sales	Sales	Tobins Q	OI/sales	Obs.
				Panel A: Entire Sample					
(1)	Fraud Profile Sim.	–0.018	0.000	0.002	0.004	0.005	0.003	–0.001	49,039
		(–3.10)	(–0.09)	(1.43)	(2.76)	(5.55)	(11.68)	(–3.71)	
(2)	Industry Similarity	0.001	–0.012	0.003	0.004	0.009	–0.001	0.000	49,039
		(0.25)	(–4.87)	(3.87)	(2.33)	(9.45)	(–2.64)	(2.87)	
				Panel B: Above Median Firm Size Only					
(3)	Fraud Profile Sim.	–0.009	0.020	0.047	0.012	0.009	0.004	–0.001	24,523
		(–1.36)	(3.52)	(2.51)	(3.61)	(4.96)	(8.65)	(–2.34)	
(4)	Industry Similarity	0.008	–0.003	–0.005	–0.002	0.004	–0.001	0.000	24,523
		(1.17)	(–0.41)	(–0.31)	(–0.52)	(2.26)	(–2.87)	(1.35)	
				Panel C: Below Median Firm Size Only					
(5)	Fraud Profile Sim.	–0.026	–0.004	0.001	0.003	0.005	0.003	–0.001	24,516
		(–2.58)	(–1.54)	(1.09)	(1.90)	(4.13)	(7.89)	(–3.00)	
(6)	Industry Similarity	–0.008	–0.017	0.004	0.005	0.011	0.000	0.000	24,516
		(–0.80)	(–6.15)	(4.00)	(2.87)	(8.84)	(–1.60)	(2.46)	
				Panel D: Entire Sample (Out of Sample Years Only)					
(7)	Fraud Profile Sim.	–0.011	–0.006	0.002	0.004	0.003	0.003	–0.001	25,926
		(–1.91)	(–1.70)	(1.33)	(1.98)	(2.21)	(4.04)	(–3.22)	
(8)	Industry Similarity	–0.005	–0.014	0.003	0.008	0.010	–0.002	0.000	25,926
		(–0.78)	(–3.58)	(2.47)	(2.73)	(5.85)	(–2.48)	(1.40)	
				Panel E: Same as Panel A, but Add Additional Controls					
(9)	Fraud Profile Sim.	–0.018	0.001	0.001	0.004	0.004	0.003	–0.001	49,039
		(–3.26)	(0.58)	(1.36)	(2.83)	(4.62)	(11.78)	(–3.54)	
(10)	Industry Similarity	0.000	–0.010	0.003	0.004	0.008	–0.001	0.001	49,039
		(0.04)	(–4.21)	(3.69)	(2.38)	(8.67)	(–2.70)	(3.06)	
				Panel F: Same as Panel A, but Replace Firm Effects with Industry Effects					
(11)	Fraud Profile Sim.	0.011	0.003	0.004	-0.003	0.001	0.006	-0.000	49,745
		(2.02)	(1.16)	(5.08)	(-2.14)	(2.97)	(14.38)	(-1.37)	0.025
(12)	Industry Similarity	0.003	-0.013	0.009	0.005	0.008	0.001	0.000	49,745
		(0.62)	(-6.27)	(11.23)	(2.65)	(19.51)	(2.48)	(2.64)	0.198

Table 8: LDA Topics Driving Fraud Similarities (Page 1 of 2)

The table lists the Topic Model Factors found to be statistically significant regarding their link to firms involved in AAER actions as compared to firms not involved in AAER actions (the last column). In addition to the topic commongrams and the representative paragraphs that describe each topic's content, the table displays coefficients and t-statistics for regressions where firm-year topic loadings are regressed on the AAER dummy. We only report results for the 5% level significant topics from the set in Table IX of Ball, Hoberg and Maksimovic (2013). These regressions include controls for firm fixed effects and year fixed effects. Hence the significant topics listed here are statistically different for firms involved in AAERs even compared to the same firm prior to or after the AAER years. The reported t-statistics are also adjusted for clustering by firm.

	Topic Commongrams	Representative Paragraph	Different in AAER years
1	partially offset, primarily due, offset decrease, due primarily, decreased decrease	Income from operations for 1997 totaled $974,549, an increase of $121,707 (14.3%) from 1996. The increase was due primarily to the increase in sales, the decreased percentage of the general and administrative expenses and the decrease in depreciation and amortization expense. Income from operations for 1996 totaled $852,842, a decrease of $40,532 (4.5%) from 1995. The decrease was due primarily to the increased costs of operating the company-owned restaurants.	-0.173 (-3.55)
2	board directors, executive officers, officers directors, vice president, directors officers	Since joining the company in January 1998, the new chief executive officer, along with the rest of the company's management team has been developing a broad operational and financial restructuring plan. A broad outline of that plan has been presented to the company's board of directors in march 1998. The plan, which is designed to leverage the company's brand, distribution and technology strengths, includes reducing costs, outsourcing of certain components and products, disposition of certain assets and capitalizing on the company's patented digital television technologies. Restructuring costs must be incurred to implement the plan.	-0.196 (-3.20)
3	legal proceedings, bankruptcy court, litigation settlement, settlement litigation, proprietary rights	Assurance that any patent owned by the company will not be invalidated, circumvented or challenged, that the rights granted thereunder will provide competitive advantages to the company or that any of the company's pending or future patent applications will be issued with the scope of the claims sought by the company, if at all. Furthermore, there can be no assurance that others will not develop similar products or software, duplicate the company's products or software or design around the patents owned by the company or that third parties will not assert intellectual property infringement claims against the company. In addition, there can be no assurance that foreign intellectual property laws will adequately protect the company's intellectual property rights abroad. The failure of the company to protect its proprietary rights could have a material adverse effect on its business, financial condition and results of operations.	-0.185 (-2.73)
4	sufficient meet, additional financing, sources liquidity, raise additional, additional funds	The company believes that its current cash, cash equivalents and short-term investment balances and cash flow from operations, if any, will be sufficient to meet the company's working capital and capital expenditure requirements for at least the next twelve months. Thereafter, the company may require additional funds to support its working capital requirements or for other purposes and may seek to raise such additional funds through public or private equity financing or from other sources. There can be no assurance that additional financing will be available at all or that if available, such financing will be obtainable on terms favorable to the company.	-0.115 (-2.63)
5	acquisition, connection acquisition, acquired businesses, completed acquisition, acquisition accounted	In august 1996, Prologic completed its acquisition of Basis. All of the outstanding stock of basis was acquired for 337,349 shares of common stock of the company plus $500,000 in cash. The acquisition was accounted for as a purchase and accordingly the purchase price and all expenses directly associated with the acquisition were allocated to the assets acquired and the liabilities assumed based on their fair market values at the date of the acquisition determined by management estimates. Goodwill in the amount of $1,459,661 was recorded in connection with the acquisition.	0.157 (2.61)
6	marketing expenses, professional fees, salaries benefits, expenses related, related expenses	Sales and marketing expenses increased to $835,000 for the year ended december 31, 1997 from $221,000 in 1996. The increase was due primarily to compensation and recruiting costs, product design costs, advertising expenses and other marketing expenses related to introduction of the company's infant jaundice product. Sales and marketing expenses are expected to increase in the future as the company begins to market this product.	-0.102 (-2.49)

Table 8: LDA Topics Driving Fraud Similarities (Page 2 of 2)

The table lists the Topic Model Factors found to be statistically significant regarding their link to firms involved in AAER actions as compared to firms not involved in AAER actions (the last column). In addition to the topic commongrams and the representative paragraphs that describe each topic's content, the table displays coefficients and t-statistics for regressions where firm-year topic loadings are regressed on the AAER dummy. We only report results for the 5% level significant topics from the set in Table IX of Ball, Hoberg and Maksimovic (2013). These regressions include controls for firm fixed effects and year fixed effects. Hence the significant topics listed here are statistically different for firms involved in AAERs even compared to the same firm prior to or after the AAER years. The reported t-statistics are also adjusted for clustering by firm.

	Topic Commongrams	Representative Paragraph	Different in AAER years
7	product line, product lines, product sales, distribution channels, product introductions	Historically, the company has introduced several new products each year. in prior years any increase in sales volume related to the new products was offset by discontinued products. In 1997, the company's product line was increased to 120 products from approximately 110 in 1996. Most of the new products were part of the new "Caribe Line" which replaced two colognes. As a result, sales of the fragrance product line increased approximately $342,000, or 163%, to $552,000 in 1997 compared to $210,000 in 1996. In 1998, the company plans to continue to expand its product line in an attempt to increase net product sales in North America.	0.086 (2.23)
8	cash flow, cash flows, cash cash equivalents, cash provided, cash investing activities	At December 31, 1997, the company had cash of $0.8 million. During the year ended December 31, 1997, the company used $27.4 million in cash which consisted of $4.8 million in net cash provided by operating activities, $60.1 million net cash used in investing activities and $28 million of net cash provided by financing activities.	-0.119 (-2.14)
9	payments made, principal payments, payment dividends, pay dividends, dividends paid	The company's license agreements currently in effect generally provide, and it is expected that future license agreements will provide, for the company to receive a payment at the time of execution of the agreement, additional scheduled payments or payments based on the attainment of certain milestones and royalty payments based on net sales of products by the licensee. The timing and amount of such payments will fluctuate, and such fluctuations could have a material adverse effect on the company's cash position and results of operations.	-0.090 (-2.06)
10	gain sale, held sale, sale leaseback, gains sale, realized gains	The gain on sale of assets of $4.2 million, recognized in 1997, was associated with the sale of the company's interest in several blocks in India, the sale of an investment in a Philippines company and the sale of the Gulf of Mexico properties. The 1996 gain on sale of assets of $1.0 million was from the sale of certain interests in India.	-0.114 (-2.03)
11	continued growth, business strategy, growth strategy, business opportunities, core business	The company's business has grown significantly since its inception, and the company anticipates future growth. the growth of the company's business and the expansion of its customer base have resulted in a corresponding growth in the demands on the company's management and personnel and its operating systems and internal controls. Any future growth may further strain existing management resources and operational, financial, human and management information systems and controls, which may not be adequate to support the company's operations.	0.126 (2.03)
12	clinical trials, research development, collaborative partners, collaborative arrangements, regulatory approvals	The company expects to incur substantial additional research and development expense including continued increases in personnel and costs related to preclinical testing and clinical trials. The company's future capital requirements will depend on many factors, including the rate of scientific progress in its research and development programs, the scope and results of preclinical testing and clinical trials, the time and costs involved in obtaining regulatory approvals, the costs involved in filing, prosecuting and enforcing patent claims, competing technological and market developments, the cost of manufacturing scale-up, commercialization activities and arrangements and other factors not within the company's control. The company intends to seek additional funding through research and development relationships with suitable potential corporate collaborators and/or through public or private financings. There can be no assurance that additional financing will be available on favorable terms, if at all.	-0.020 (-2.02)

Table 9: LDA Topics Driving Fraud Similarities (placebo tests)

The table lists the Topic Model Factors found to be statistically significant regarding their link to firms involved in AAER actions as compared to firms not involved in AAER actions (first column after Topic Descriptions). We also report significant topics for firms in the year after, and also the year before, they are alleged to be involved in fraud (last two columns). The table displays coefficients and t-statistics for regressions where firm-year topic loadings are regressed on the AAER dummy, the post-AAER dummy, and the pre-AAER dummy, respectively. We only report results for the 5% level significant topics from the set in Table IX of Ball, Hoberg and Maksimovic (2013). These regressions include controls for firm fixed effects and year fixed effects. Hence the significant topics listed here are statistically different for firms involved in AAERs even compared to the same firm prior to or after the AAER years. The reported t-statistics are also adjusted for clustering by firm.

			Different in	Different in
		Different in	Pre-	Post-
	Topic Commongrams	AAER years	AAER years	AAER years
1	partially offset, primarily due, offset decrease, due primarily, decreased decrease	-0.173 (-3.55)		
2	board directors, executive officers, officers directors, vice president, directors officers	-0.196 (-3.20)		
3	legal proceedings, bankruptcy court, litigation settlement, settlement litigation, proprietary rights	-0.185 (-2.73)	-0.172 (-2.32)	0.222 (2.89)
4	sufficient meet, additional financing, sources liquidity, raise additional, additional funds	-0.115 (-2.63)		
5	acquisition, connection acquisition, acquired businesses, completed acquisition, acquisition accounted	0.157 (2.61)		-0.134 (-2.40)
6	marketing expenses, professional fees, salaries benefits, expenses related, related expenses	-0.102 (-2.49)		0.107 (2.25)
7	product line, product lines, product sales, distribution channels, product introductions	0.086 (2.23)		
8	cash flow, cash flows, cash cash equivalents, cash provided, cash investing activities	-0.119 (-2.14)		
9	payments made, principal payments, payment dividends, pay dividends, dividends paid	-0.090 (-2.06)		0.149 (2.63)
10	gain sale, held sale, sale leaseback, gains sale, realized gains	-0.114 (-2.03)		
11	continued growth, business strategy, growth strategy, business opportunities, core business	0.126 (2.03)		
12	clinical trials, research development, collaborative partners, collaborative arrangements, regulatory approvals	-0.020 (-2.02)		
13	license fees, consulting services, consulting fees, service fees, services provided		-0.145 (-3.10)	
14	past years, recent years, significant portion, substantial portion, years company		-0.160 (-2.95)	
15	senior notes, principal amount, notes payable, subordinated notes, senior subordinated notes		0.219 (2.76)	
16	laws regulations, government regulation, federal state, government agencies, change control		-0.158 (-2.52)	
17	generally accepted, conducted audits accordance generally accepted auditing, based, principles significant estimates made management		-0.138 (-2.39)	
18	life insurance, premiums written, insurance premiums, premiums earned, insurance coverage		-0.119 (-2.35)	
19	foreign currency, foreign exchange, north america, currency exchange, domestic international		0.130 (2.02)	-0.103 (-2.08)
20	restructuring charge, restructuring charges, write downs, special charges, fourth quarter			0.259 (3.68)
21	interest rates, certificates deposit, asset liability, assets liabilities, balance sheet			-0.115 (-2.03)
22	entered agreement, agreement dated, terms agreement, pursuant terms, agreement entered			0.111 (2.02)

Table 10: LDA Topics Driving Fraud Similarities (Revenue and Expense Fraud)

The table lists the Topic Model Factors found to be statistically significant regarding their link to firms involved in specific types of AAER actions (revenue and expense fraud) as compared to firms not involved in these actions. We thus report significant topics separately for firms involved in revenue fraud and expense fraud. In addition to the topic commongrams and the representative paragraphs that describe each topic's content, the table displays coefficients and *t*-statistics for regressions where firm-year topic loadings are regressed on the Revenue-AAER dummy and the Expense-AAER dummy, respectively. We only report results for the 5% level significant topics from the set in Table IX of Ball, Hoberg and Maksimovic (2013). These regressions include controls for firm fixed effects and year fixed effects. Hence the significant topics listed here are statistically different for firms involved in AAERs even compared to the same firm prior to or after the AAER years. The reported *t*-statistics are also adjusted for clustering by firm.

	Topic Commongrams	Representative Paragraph	Revenue AAER years	Expense AAER years
1	partially offset, primarily due, offset decrease, due primarily, decreased decrease	Income from operations for 1997 totaled $974,549, an increase of $121,707 (14.3%) from 1996. The increase was due primarily to the increase in sales, the decreased percentage of the general and administrative expenses and the decrease in depreciation and amortization expense. Income from operations for 1996 totaled $852,842, a decrease of $40,532 (4.5%) from 1995. The decrease was due primarily to the increased costs of operating the company-owned restaurants.	-0.221 (-2.57)	-0.239 (-2.42)
2	sufficient meet, additional financing, sources liquidity, raise additional, additional funds	The company believes that its current cash, cash equivalents and short-term investment balances and cash flow from operations, if any, will be sufficient to meet the company's working capital and capital expenditure requirements for at least the next twelve months. Thereafter, the company may require additional funds to support its working capital requirements or for other purposes and may seek to raise such additional funds through public or private equity financing or from other sources. There can be no assurance that additional financing will be available at all or that if available, such financing will be obtainable on terms favorable to the company.	-0.162 (-2.60)	
3	total revenues, revenues derived, percentage revenues, revenues revenues, revenues generated	Revenues increased by $29.9 million, or approximately 27.4%, to $139.1 million in 1997 from $109.2 million in 1996. This increase was primarily due to revenues generated by increased sales of commercial airtime inventory. acquisitions accounted for $5.7 million of this increase. excluding these revenues, same market revenues increased $24.2 million in 1997, or 22.2%. Revenues from reciprocal arrangements as a percentage of total revenues declined to 4.0% in 1997 from 8.0% in 1996.	0.162 (2.33)	
4	partially offset, offset lower, partially offset lower, due higher, due lower	Earnings in 1996 were $34.3 million better than 1995. The improvement was mainly due to higher volumes in the company's major product lines, higher cement and U.S. ready-mixed concrete prices, lower operating costs in construction materials operations and lower imports to supplement production in the U.S. these increases were partially offset by lower divestments gains and lower clinker production in Canada.	-0.138 (-2.20)	
5	clinical trials, research development, collaborative partners, collaborative arrangements, regulatory approvals	The company expects to incur substantial additional research and development expense including continued increases in personnel and costs related to preclinical testing and clinical trials. The company's future capital requirements will depend on many factors, including the rate of scientific progress in its research and development programs, the scope and results of preclinical testing and clinical trials, the time and costs involved in obtaining regulatory approvals, the costs involved in filing, prosecuting and enforcing patent claims, competing technological and market developments, the cost of manufacturing scale-up, commercialization activities and arrangements and other factors not within the company's control. The company intends to seek additional funding through research and development relationships with suitable potential corporate collaborators and/or through public or private financings. There can be no assurance that additional financing will be available on favorable terms, if at all.	-0.039 (-2.14)	
6	legal proceedings, bankruptcy court, litigation settlement, settlement litigation, proprietary rights	Assurance that any patent owned by the company will not be invalidated, circumvented or challenged, that the rights granted thereunder will provide competitive advantages to the company or that any of the company's pending or future patent applications will be issued with the scope of the claims sought by the company, if at all. Furthermore, there can be no assurance that others will not develop similar products or software, duplicate the company's products or software or design around the patents owned by the company or that third parties will not assert intellectual property infringement claims against the company. In addition, there can be no assurance that foreign intellectual property laws will adequately protect the company's intellectual property rights abroad. The failure of the company to protect its proprietary rights could have a material adverse effect on its business, financial condition and results of operations.	-0.210 (-2.06)	
7	research development, research development expenses, product development, process research development, development stage	Research and development expenses increased 20.7% to $6,006,000 in 1996, and increased as a percentage of net sales to 10.0% in 1996 from 6.1% in 1995. The increases in research and development expenses were primarily due to the expansion of the research and development staff, and expenses associated with its research and development facility. The majority of these increased costs related to the support of the apex product development.		0.307 (2.49)
8	product line, product lines, product sales, distribution channels, product introductions	Historically, the company has introduced several new products each year. in prior years any increase in sales volume related to the new products was offset by discontinued products. In 1997, the company's product line was increased to 120 products from approximately 110 in 1996. Most of the new products were part of the new "Caribe Line" which replaced two colognes. As a result, sales of the fragrance product line increased approximately $342,000, or 163%, to $552,000 in 1997 compared to $210,000 in 1996. In 1998, the company plans to continue to expand its product line in an attempt to increase net product sales in North America.		0.103 (2.49)
9	management believes, management team, asset management, management aware, independent auditors	Pursuant to the management services agreement with standard management, premier life (luxembourg) paid standard management a management fee of $25,000 per quarter during 1997 and 1996 for certain management and administrative services. The agreement provides that it may be modified or terminated by either standard management or premier life (luxembourg).		0.206 (2.28)
10	efforts reduce, order reduce, effort reduce, economies scale, cutting measures	Factors which management believes may affect the future financial performance of the company include but are not limited to: the successful implementation of manufacturing and business processes which will reduce costs and improve efficiency; the investment in engineering and marketing activities which lead to improved sales growth; the successful integration of acquisitions into the company's operations; dealing with the external regulatory influences on the company's primary markets; and the effect on the competitive environment resulting in the consolidation of companies within the instrumentation industry.		-0.359 (-2.21)

Table 11: Fog Index Regressions

The table reports OLS regressions for our sample of observations based on annual firm observations from 1997 to 2008. One observation is one firm in one year. The dependent variable is a fog index or readability index as noted in the first column. All three readability indices are constructed such that a higher value indicates greater difficulty in reading. Panels A to C differ in how the AAER Dummy is lagged. The AAER dummy in Panel A is one if an AAER action indicates that the firm was involved in fraudulent activity in the current year. In Panel B, the AAER dummy is one in the year prior to a year in which a given firm was involved in an AAER, and In Panel C, the AAER dummy is one in the year after a given firm was involved in an AAER. See Table 1 for the description of our key variables. All regressions are estimated with year and firm fixed effects, and standard errors are clustered by firm. *t*-statistics are in parentheses.

							MD&A	MD&A	
			Operating				Peer	Peer	
	Dependent	AAER	Income	R&D	CAPX	Log	Implied	Implied	
Row	Variable	Dummy	/Sales	/Sales	/Sales	Sales	Tobins Q	OI/sales	Obs.
				Panel A: AAER-year Results					
(1)	Automated Read. Index	–0.128	–0.276	0.022	0.054	0.083	–0.013	–0.004	49,039
		(–1.74)	(–6.01)	(1.42)	(1.94)	(4.46)	(–3.48)	(–1.20)	
(2)	Gunning Index	–0.183	–0.157	0.004	0.050	0.014	–0.011	–0.001	49,039
		(–2.88)	(–3.97)	(0.29)	(2.02)	(0.91)	(–3.50)	(–0.26)	
(3)	Flesch Kinkaid Index	–0.288	–0.628	0.093	0.302	0.293	–0.010	0.009	49,039
		(–0.97)	(–3.91)	(1.77)	(3.05)	(4.44)	(–0.80)	(0.72)	
				Panel B: pre-AAER-year Results					
(4)	Automated Read. Index	–0.120	–0.275	0.022	0.054	0.082	–0.013	–0.004	49,039
		(–1.08)	(–5.99)	(1.41)	(1.92)	(4.39)	(–3.51)	(–1.18)	
(5)	Gunning Index	–0.047	–0.156	0.003	0.050	0.013	–0.012	–0.001	49,039
		(–0.47)	(–3.95)	(0.28)	(1.99)	(0.80)	(–3.54)	(–0.24)	
(6)	Flesch Kinkaid Index	0.275	–0.627	0.093	0.301	0.291	–0.010	0.009	49,039
		(0.69)	(–3.91)	(1.76)	(3.04)	(4.40)	(–0.82)	(0.73)	
				Panel C: Post-AAER-year Results					
(7)	Automated Read. Index	0.241	–0.275	0.022	0.054	0.082	–0.013	–0.004	49,039
		(2.68)	(–5.98)	(1.41)	(1.93)	(4.39)	(–3.49)	(–1.20)	
(8)	Gunning Index	0.211	–0.156	0.003	0.050	0.013	–0.012	–0.001	49,039
		(2.72)	(–3.93)	(0.27)	(2.00)	(0.79)	(–3.52)	(–0.25)	
(9)	Flesch Kinkaid Index	0.689	–0.624	0.093	0.301	0.290	–0.010	0.009	49,039
		(2.26)	(–3.89)	(1.76)	(3.05)	(4.39)	(–0.81)	(0.72)	

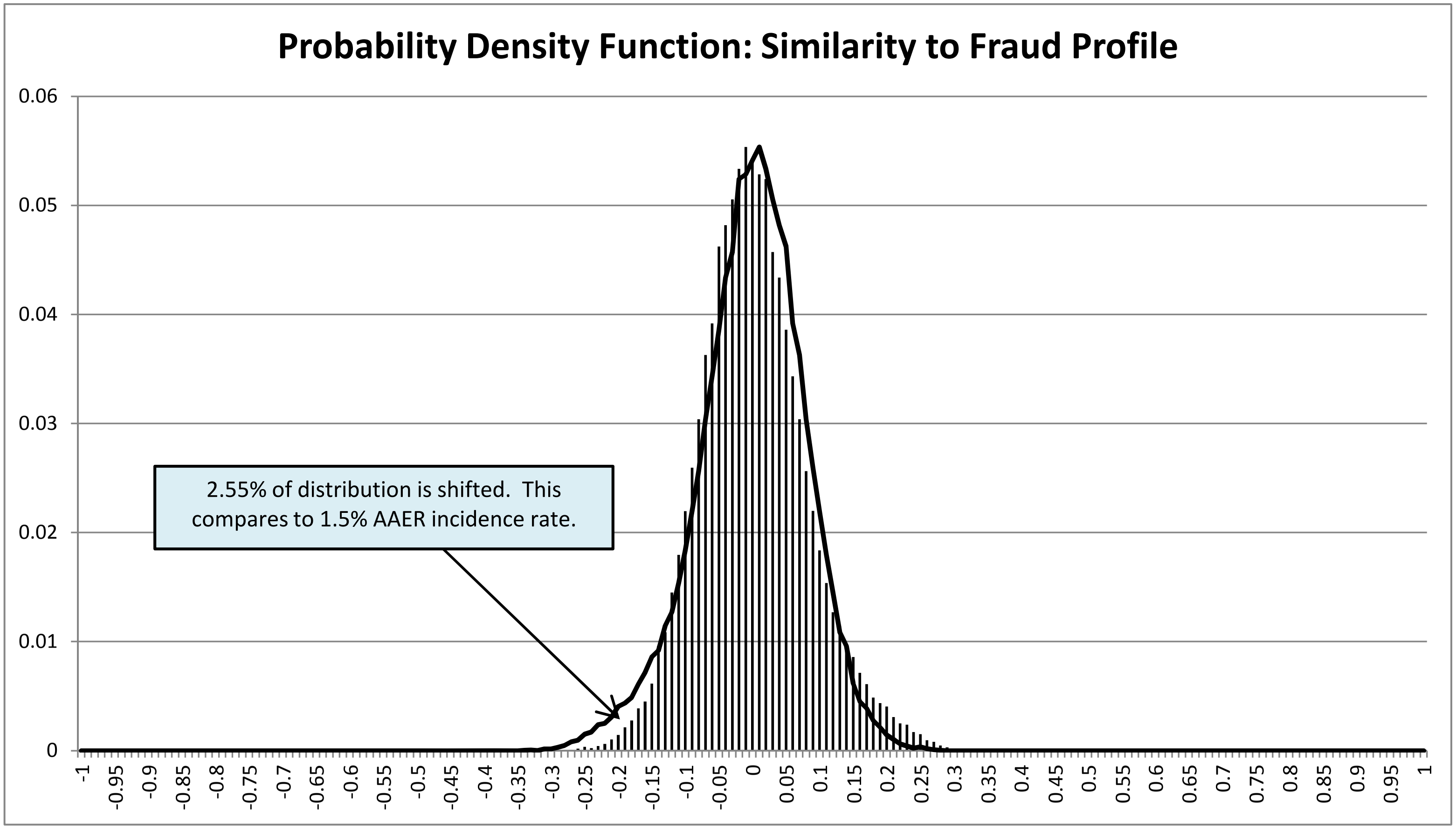
Table 12: Equity Market Liquidity and Issuance

The table reports OLS regressions for our sample of observations based on annual firm observations from 1997 to 2008. One observation is one firm in one year. The dependent variable is the fraud profile similarity, the fraud dummy, or Compustat equity issuance divided by assets as noted in the first column. All regressions include firm and year fixed effects. In Panel A, the dependent variable is either the Fraud Score or the AAER Dummy as noted in the first column. The fraud similarity score is the cosine similarity of the given firm's abnormal disclosure and the average abnormal disclosure of all firms involved in AAERs in the sample period 1997 to 2001. The AAER dummy is one in the year prior to a year in which a given firm was involved in an AAER. Equity issuance in Panel B is either Compustat equity issuance or SDC Platinum public SEO issuance. Both are in dollars and are scaled by assets. See Table 1 for the description of our key variables. All standard errors are clustered by firm. *t*-statistics are in parentheses.

		Forced					MD&A	MD&A	
		Mutual	Operating				Peer	Peer	
	Dependent	Fund	Income	R&D	CAPX	Log	Implied	Implied	
Row	Variable	Selling	/Sales	/Sales	/Sales	Sales	Tobins Q	OI/sales	Obs.
				Panel A: Firm and Year Fixed Effects					
(3)	Fraud Score	0.003	0.000	0.001	0.005	0.006	0.003	-0.001	30,683
		(5.16)	(0.11)	(0.83)	(2.18)	(4.09)	(10.38)	(-2.96)	0.012
(4)	AAER Dummy	0.005	-0.010	0.001	0.000	0.007	0.001	-0.000	30,683
		(3.97)	(-1.47)	(0.48)	(0.16)	(1.97)	(1.74)	(-0.08)	0.002

							MD&A	MD&A	
		Fraud	Operating				Peer	Peer	
	Dependent	Profile	Income	R&D	CAPX	Log	Implied	Implied	
Row	Variable	Similarity	/Sales	/Sales	/Sales	Sales	Tobins Q	OI/sales	Obs.
				Panel B: Equity Issuance: Firm and Year Fixed Effects					
(5)	Compustat Equity Issuance	0.085	-0.040	0.005	0.033	-0.037	0.017	-0.004	30,683
		(4.59)	(-3.45)	(1.07)	(4.32)	(-10.18)	(8.67)	(-3.90)	0.064
(6)	SDC Public SEO Issuance	0.051	0.017	0.002	0.018	-0.006	0.005	-0.000	30,683
		(4.67)	(2.52)	(0.74)	(3.59)	(-3.06)	(7.66)	(-0.41)	0.012

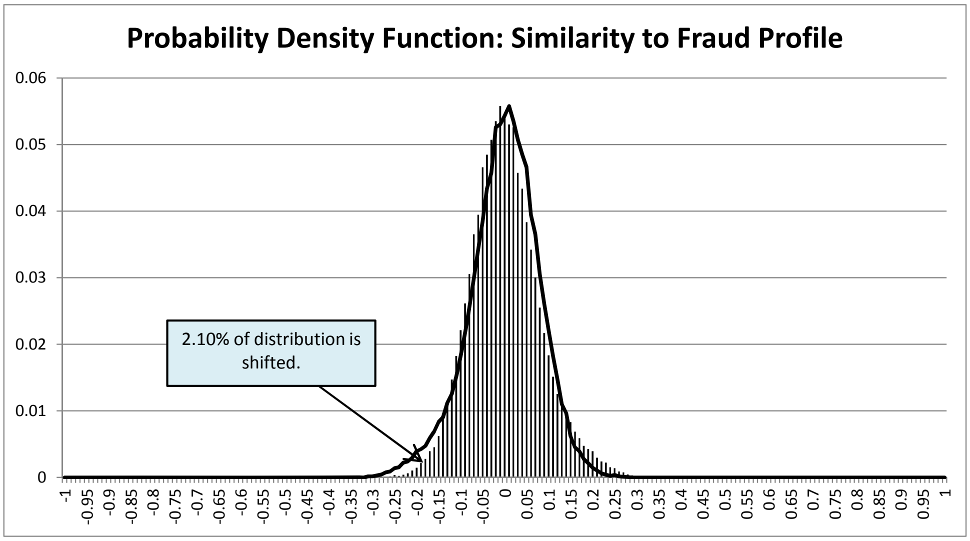
Figure 1: Empirical distribution of firm Fraud Similarities. The distribution is based on our entire sample including both firms that were involved in AAERs and firms that were not. The actual distribution is displayed using the bar chart format. To illustrate the degree of left-right asymmetry, the line plot displays the shape of the y-axis reflection of the actual distribution. The size of the asymmetric mass is then summarized.



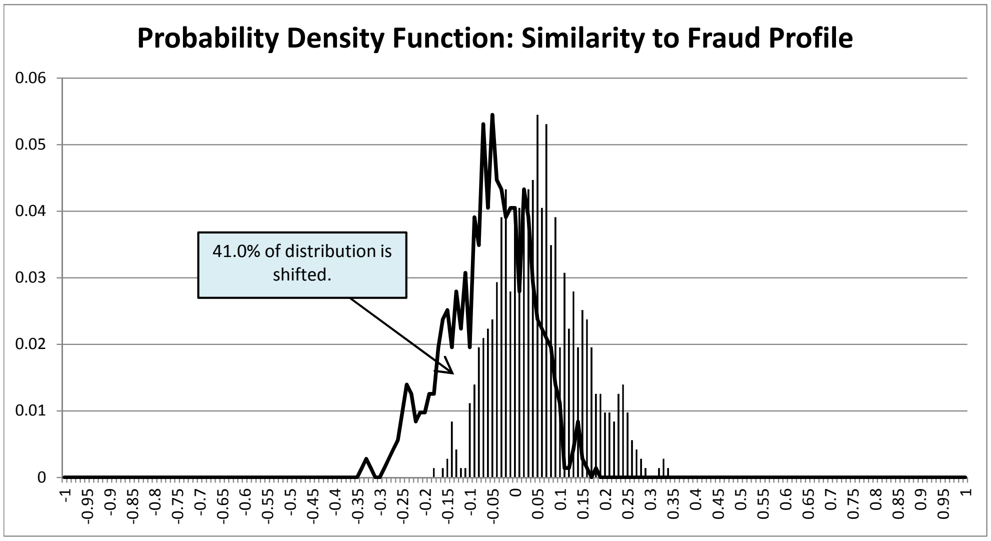
Probability Density Function: Similarity to Fraud Profile

2.55% of distribution is shifted. This compares to 1.5% AAER incidence rate.

Figure 2: Empirical distribution of firm Fraud Similarities for two subsamples. The upper figure's distribution is based on all firms in our sample excluding firm years involved in AAERs. The lower figure reports the fraud similarity distribution only for firms-years involved in AAERs. In both figures, the actual distribution is displayed using the bar chart format. To illustrate the degree of left-right asymmetry, the line plot displays the shape of the y-axis reflection of the actual distribution. The size of the asymmetric mass is then summarized.

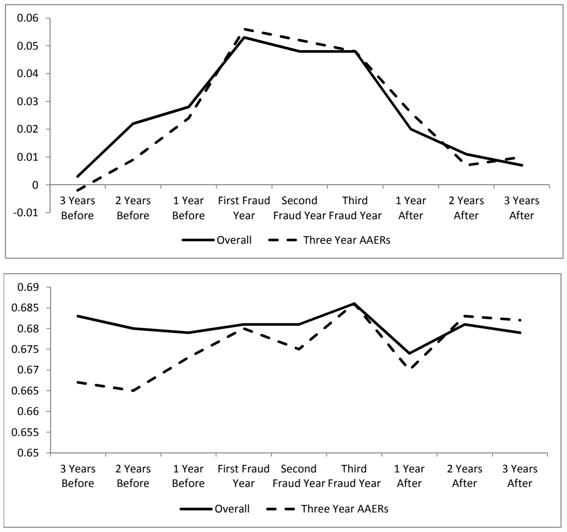


2.10% of distribution is shifted.



41.0% of distribution is shifted.

Figure 3: Average Fraud Similarities over time for firms involved in AAERs. The figure displays the average fraud similarity score during the period of time that the AAER alleges fraud occurred, and also during the period of time preceding and after the period of the alleged fraud. Regardless of duration of the fraudulent period, we tag the three years prior to the fraud period as the ex-ante period and the three years after the fraud period as the ex-post period. For firms that had a fraud period of one or two years, they would be counted in the first fraud year and the second fraud year calculation, but not the third fraud year calculation. To ensure that fraud duration is not overly influencing our results, we also display results where we limit the sample to firms with alleged fraud that lasted at least three years.


3 Years Before
2 Years Before
1 Year Before
First Fraud Year
Second Fraud Year
Third Fraud Year
1 Year After
2 Years After
3 Years After
Overall
Three Year AAERs
3 Years Before
2 Years Before
1 Year Before
First Fraud Year
Second Fraud Year
Third Fraud Year
1 Year After
2 Years After
3 Years After
Overall
Three Year AAERs