Automatic summaries of earnings releases: Attributes and effects on investors' judgments

Eddy Cardinaels
Department of Accountancy, Finance and Insurance, KU Leuven
Department of Accountancy, Tilburg University
eddy.cardinaels@kuleuven.be

Stephan Hollander
Department of Accountancy
Tilburg University
s.hollander@tilburguniversity.edu

Brian J. White
Department of Accounting
The University of Texas at Austin
brian.white@mccombs.utexas.edu

September 30, 2018

We thank Sarah Zechman (discussant), David Veenman, and workshop participants at the London Business School Accounting Symposium, the Financial Accounting and Reporting Section (FARS) Midyear Meeting, Radboud University Nijmegen, the University of Texas at Austin, the University of Melbourne, the University of Pittsburgh, the University of Amsterdam, the University of New South Wales, Jinan University, Tilburg University, NHH Norwegian School of Economics, and Texas A&M University for helpful comments. We also thank Arnaud Nicolas for generating expert summaries of earnings releases, and Ties de Kok and Shannon Garavaglia for their capable research assistance. This paper won the 2018 FARS Midyear Meeting Best Paper Award.

Automatic summaries of earnings releases: Attributes and effects on investors' judgments

ABSTRACT

Firms often include summaries with earnings releases. However, manager-generated summaries may be prone to strategic tone and content management compared to the disclosures they summarize. In contrast, computer algorithms can summarize large amounts of text without human intervention, and may provide useful summary information with less bias. We use multiple methods to provide evidence regarding the characteristics of automatic, algorithm-based summaries of earnings releases compared to summaries provided by managers. Results suggest that automatic summaries are generally less positively biased than management summaries, often without sacrificing the extent to which the summaries capture relevant information. We then conduct an experiment to test whether these differing attributes of automatic and management summaries affect individual investors' judgments. We find that investors who receive an earnings release accompanied by an automatic summary arrive at more conservative (i.e., lower) valuation judgments, and are more confident in those judgments. Overall, our results suggest that summaries affect investors' judgments, and that such effects can differ for management and automatic summaries.

Keywords: *Management summary; automatic summary; corporate disclosure; individual investors; investor judgment*

Data availability: *Contact the authors*

1. Introduction

Public companies disclose more information than ever before (e.g., KPMG 2011; Loughran and McDonald 2014; Dyer et al. 2017). Given the large volume of disclosure and evidence that individual investors are boundedly rational (Hirshleifer and Teoh 2003; Elliott et al. 2015), firms often provide summaries of key disclosures, including earnings releases. However, rather than presenting a balanced picture of the information disclosed in the underlying document, managers may engage in tone management and/or selectively highlight information that is favorable to the company (Henry 2008; Guillamon-Saorin et al. 2012; Huang et al. 2013, 2014). Against this backdrop, there may be a role for automatic, algorithm-based summarization of earnings releases. Summarization algorithms rely on statistical heuristics for sentence extraction to summarize large amounts of text without human intervention. As such, automatic summaries have the potential to reduce both information overload and bias. In this study, we investigate two questions. First, how do management and automatic summaries of earnings releases compare on a range of attributes, including bias and the extent to which they capture important information in the earnings release? Second, how do automatic and management summaries of earnings releases affect individual investors' judgments?

Investigating these issues is important for several reasons. First, because disclosures have become lengthier (Dyer et al. 2017), regulators and standard setters are starting to explore ways of simplifying financial reports (SEC 2013; FASB 2015), including summarization (SEC 2016). These efforts have led to calls for research on summarization to aid investors and others (Barth 2015). Thus, investigating management and automatic summaries has the potential to provide new insights to financial reporting regulators and accounting standard setters. Second, manager-generated summaries are already part of the financial reporting landscape. Our review of S&P

100 firms' disclosure practices indicates that 81% provided summaries in their fourth quarter 2015 earnings releases. However, there is scant evidence on the attributes of these summaries or whether this practice affects investors' judgments.

We use multiple methods to address our research questions. For our first research question, we start by selecting a summarization algorithm, of which there are several that are widely available (Nenkova and McKeown 2011). To identify an algorithm that is well-suited to summarizing earnings releases, we rely on user evaluation, and ask participants from Amazon's Mechanical Turk (MTurk) platform to evaluate summaries of two earnings releases: one summary provided by management and others generated by computer algorithms. To validate these evaluations, we use a technique from the field of information retrieval, known as Recall-Oriented Understudy for Gisting Evaluation (ROUGE), in which summaries are evaluated against a summary prepared by an experienced Investment Relations Officer (IRO). Based on these analyses, we select a summarization algorithm known as LexRank and compare the tone of management and LexRank summaries for a larger sample of hand-collected earnings releases.[1] To address our second research question, we conduct an experiment to investigate the effect of automatic and management summaries on investors' valuation and other investment-related judgments.

In our user evaluation test, MTurk participants compare automatic and management summaries to the underlying text of two earnings releases on several dimensions. Results suggest that automatic summaries reflect the underlying text of the earnings release with less bias (i.e., present a more balanced picture) than management summaries. Participants also rate automatic summaries as no different from management summaries in capturing the important information

[1] LexRank is short for *Lexical PageRank*, a method developed by Larry Page, one of the founders of Google.

in the earnings releases, and participants are equally likely to rely on automatic and management summaries. Comparing the various summarization algorithms, we find that LexRank outperforms the other algorithms, by producing summaries that are consistently rated as equal or superior to management summaries.[2] This finding is supported by the ROUGE evaluation: compared to the management summary and other automatic summaries, LexRank better captures elements of the earnings release that the IRO deems important.

We then employ the LexRank algorithm to generate automatic summaries for a hand-collected sample of S&P 100 firms that provide summaries with their fourth quarter 2015 earnings releases. We compare the tone of these automatic summaries to the tone of the management summaries, and compare the tone of both summary types to the tone of the underlying text in the earnings release. Results indicate that management summaries contain more positive tone words, and fewer negative tone words, than automatic summaries. Further, management summaries are more positive, and less negative, in tone compared to the underlying text in the earnings release, whereas automatic summaries are more similar in tone to the underlying text.

Collectively, these results, obtained with different methods, suggest that management summaries of earnings releases exhibit incremental, positive bias compared to the underlying text of the earnings release, but that automatic summaries largely do not exhibit such bias. Notably, our user evaluation and ROUGE tests suggest that this reduction in bias can be achieved without loss of informativeness.

In our main analysis, we experimentally test whether the differing attributes of automatic and management summaries affect individual investors' valuation and other investment-related

[2] We provide detail on the summarization algorithms used in this study in the Online Appendix. For further details on the LexRank algorithm specifically, we refer the reader to Section 2.1.

judgments. Although we used actual disclosures of real companies in our earlier tests, for the experiment we follow prior literature (e.g., Rennnekamp 2012) and anonymize an earnings release from a real company so that participants' familiarity with the real company does not affect their judgments. In the experiment, MTurk participants receive an anonymized earnings release that includes either a management summary, an automatic summary, or no summary, resulting in a 1 × 3 between-participants design. We generate the automatic summary using LexRank, and we confirm that the automatic and management summaries differ in the respects suggested by our earlier tests. All participants also receive linked access to the full text of the earnings release and accompanying tables, which they can search for additional information.

Our experimental results indicate that participants who receive the earnings release with the management summary place a significantly higher value on the company's stock than those who receive the automatic summary or no summary. Importantly, however, participants who receive the automatic summary are more confident in their (lower) valuation judgments than those who receive the management summary. Our results further show that judgments of future earnings growth potential and perceptions of the favorability of the earnings release explain the effect of summary type (i.e., management vs. automatic) on common stock valuation.

This paper makes several contributions. First, our paper is among the first to examine text summarization techniques in the context of earnings releases and their effects on investors' judgments. In so doing, our paper makes an incremental contribution to research on the impact of computer-based textual analysis and linguistic processing technology on investor behavior (see Loughran and McDonald 2016 for a review). Specifically, our results indicate both that managers tend to bias their summaries beyond any bias present in the underlying disclosure, and that the use of technology to generate automatic summaries can potentially avoid this bias. For

regulators and standard setters interested in summary information, our study provides evidence that encouraging management-generated summaries would not necessarily lead to the most value-relevant information being highlighted in an unbiased way.

Second, we contribute to the literature, spanning accounting, economics and finance, on individual investors' bounded rationality. Analytical and empirical studies find that bounded rationality—i.e., individual investors displaying limited attention and processing power—affects market price efficiency (e.g., Daniel et al. 2002; Hirshleifer and Teoh 2003; Hirshleifer et al. 2009; Elliott et al. 2015; Umar 2017). Further, regulators have expressed concerns that in the presence of bounded rationality, information overload can exacerbate inefficiency (Paredes 2003, 2013). Providing summarized information may be considered an easy fix for this problem. However, we find that individual investors' reactions to summary information depend on whether the summary is generated automatically or by management.

Third, while information intermediaries (e.g., analysts, business journalists) contribute to the efficient allocation of capital, research also consistently shows that conflicts of interest (e.g., *quid pro quo* relations between journalists and their sources, analysts' incentives to collude with management; Desai et al. 2016, Dyck and Zingales 2003) and behavioral biases tend to stand in the way of information intermediaries fulfilling their potential. Our paper contributes to this literature by highlighting the possibility of automating one aspect of information dissemination in capital markets, thereby removing one opportunity for conflicts of interest and behavioral biases to negatively affect information.

Finally, in a recent paper, Blankespoor et al. (2018) examine the capital markets effects of news articles written by Associated Press (AP) algorithms about firms' earnings announcements. They find that these "robo-journalism" articles increase liquidity and trading volume for firms

covered by these articles, and that these effects are likely due to increased attention and awareness by retail investors. Our research is related to, but distinct from, Blankespoor et al. (2018) in several important ways. First, the technology we examine differs from that used in Blankespoor et al. (2018). AP's algorithms identify pre-selected items (EPS, etc.) from earnings releases and other data sources, and then use Natural Language Generation (NLG) algorithms to write an article using language that differs from the language used in the original information sources. Specifically, the articles contain only basic facts without interpretation or spin. In contrast, the sentence-extraction-based methods that we examine produce summaries without changing the language used in the underlying text of the earnings release. Second, whereas Blankespoor et al. (2018) focus solely on algorithmically generated articles, we compare the attributes of algorithm-generated summaries to summaries generated by managers. We argue (and find) that summarization tools can capture relevant information in earnings releases with less bias than management summaries. Third, in contrast to Blankespoor et al. (2018), who examine volume and liquidity effects from AP's automated articles, we conduct an experiment to provide causal evidence of the valuation effects of automatic and management summaries.

The paper proceeds as follows. Section 2 provides background for our study. Section 3 investigates attributes of management and automatic summaries. Section 4 develops hypotheses and presents the design and results of our experiment testing the effects of summaries on investors' judgments. Section 5 concludes the paper.

2. Background

As corporate disclosures get longer (Dyer et al. 2017; Francis et al. 2002), individual investors with limited attention may find it difficult to process all information contained in company disclosures. This suggests that it may be useful to study how summaries help (or

hinder) individual investors in making investment-related judgments, an issue that policy makers also deem to be relevant (SEC 2013, 2016). The question of how summaries affect individual investors' investment-related judgments is even more important when one considers the flexibility that the Securities and Exchange Commission (SEC) offers to its registrants, as the SEC does not offer any guidance on summary length or the items that a summary should cover (SEC 2016, 3-4).[3]

Despite the recognition that summarization can be useful in the domain of corporate disclosures, there has been no systematic research on how summarization affects individual investors' judgments (Barth 2015). Our paper provides evidence on this important issue in two ways. First, we test the viability of using automatic summarization techniques in practice by generating automatic summaries for real companies' earnings releases and comparing them to management summaries on several dimensions (section 3). Second, we test the impact of summaries on investors' judgments (section 4).

In the remainder of this section, we provide background on the technology underlying automatic summarization, discussing differences in how management and automatic summaries are generated (section 2.1), and how such differences may affect information coverage and linguistic tone (section 2.2).

2.1 Extraction-Based Automatic Summarization

The Online Appendix provides a primer on extraction-based automatic summarization, including details on six widely-available algorithms. These algorithms differ principally in the

[3] See Henry (2008), for the regulatory context. In June 2016, the SEC adopted an interim final rule that allows issuers to include, at their option, a summary page in Form 10-K. As noted therein, summary information must be presented "fairly and accurately."

statistical heuristics applied to identify the most salient sentences of a document.[4] As will become clear later, the algorithm that performs best on earnings releases is LexRank (Erkan and Radev 2004). To give the reader a sense of how summarization algorithms work in general, and how LexRank works specifically, we briefly describe the LexRank algorithm here.

The LexRank algorithm first generates a graph, composed of all sentences in the underlying document. In this graph, each sentence represents a node, with similarity between nodes as edges. To define similarity, each sentence is represented as a "bag-of-words" model, meaning that grammar and the order of words in a sentence are disregarded. The similarity between two sentences is computed by the frequency of word occurrence (specifically, tf*idf cosine similarity) in a sentence. A similarity matrix is then constructed, wherein each entry is the similarity between a sentence pair. Based on the concept of eigenvector centrality, the LexRank algorithm computes centrality-based sentence salience: sentences that are similar to many of the other sentences are considered more salient—or central—to the (unspecified) theme of the document (Erkan and Radev 2004; Newman 2008). Finally, a summary is generated by combining the top-ranked sentences, using a threshold and length cutoff to limit the size of the summary. In our implementation, we set the number of sentences per summary equal to the number of bullet points in management's summary.

2.2 Human versus Algorithm-Based Summarization

Human-generated summaries are typically based on text understanding (i.e., summarization by abstraction). A typical process for a person generating a summary would involve (1) getting an understanding of the content of the source document, (2) identifying the most relevant content

[4] In our tests, we employ a variety of frequently used sentence-extraction-based approaches for generic summarization, applicable for settings in which no additional information or prior knowledge (e.g., about user need) is needed. We exclude from our analysis genre-specific (e.g., academic journal articles) and domain-specific (e.g., medical) approaches; see Nenkova and McKeown (2011) for an overview.

contained in the document, and (3) writing up this information, usually in the person's own words (Brandow et. al. 1995). Importantly, steps 1 and 3 are beyond the capability of most automatic summarization techniques (Brandow et al. 1995; Salton et al. 1997; Nenkova and McKeown 2011). For step 2, as explained above, automatic summarization algorithms rely on statistical heuristics that attempt to identify the most important lexical units (typically, sentences) in a document (i.e., summarization by extraction).

Given that management should have a good understanding of the information content of the underlying source document (step 1), one would expect that with regard to content selection (step 2), management should be able to highlight information that investors deem relevant. However, ample evidence suggests managers tend to bias disclosures when they have discretion to do so. Given that the set of news items in the underlying document tends to be large (at least for companies with reasonably complex operations; Henry 2008), we expect managers to select items that depict a more favorable view of the company's performance when they write up a summary. That is, we expect *incremental* bias in management summaries compared to any bias already present in the underlying document. Consistent with this, Henry (2008, 371) describes selective inclusion of information in bulleted introductory points of earnings releases (e.g., a firm disclosing an increase in operating margin for one of its divisions, while overall profit margin of the company declined). Other studies (e.g., Ahern and Sosyura 2014; Guillamon-Saorin et al. 2012) find similar evidence of content management by managers. In contrast to a management summary, an extraction-based summarization algorithm selects lexical units based on statistical heuristics. If a sentence is deemed important according to the statistical heuristics, the automatic summary will include it regardless of whether it is good or bad news.

Regarding write-up (step 3), in addition to content selection, managers can also rewrite and manage the tone of the content that they include in a summary. Prior research shows that managers use tone to depict a more positive view of the company, for example by using positive tone words and vivid language (e.g., Davis and Tama-Sweet 2012; Hales et al. 2011; Henry 2008; Henry and Leone 2016; Huang et al. 2013; Loughran and McDonald 2016). Thus, we expect that managers manage tone when preparing summaries, given that they have incentives and discretion to do so (Arslan-Ayaydin et al. 2016). In contrast, extraction-based summarization algorithms cannot change the language that appears in the summary given that they extract sentences directly from the source document. In other words, the algorithm is more likely to reflect the tone of the underlying release.[5]

The next section explicitly tests these presumptions by comparing attributes of automatic and management summaries.

3. Attributes of Automatic and Management Summaries

In this section, we compare the attributes of automatic and management summaries of earnings releases on several dimensions. To accomplish this, we first use validated techniques and methods from the communications and information retrieval literature, and apply them to two real earnings releases (Loughran and McDonald 2016). Then, we conduct archival analysis of differences in tone between automatic and management summaries for a larger sample of hand-collected earnings releases.

[5] While management has incentives to positively bias summaries (e.g. by downplaying bad news or adding positive tone), it is difficult to define what optimal tone should be. For example, management might use more negative tone in the underlying earnings release in order to reduce litigation risk. If this negative tone is reflected in an automatic summary, then automatic summaries may be negatively biased. Thus, although we expect automatic summaries to be less positive than management summaries, and to better reflect the content and tone of the underlying text, we stop short of predicting that either summary type will be more or less optimal.

3.1 User Evaluation

3.1.1 Overview and Participants

Following recent studies in communications research, in which users assess algorithm- versus human-generated news content (e.g., Clerwall 2014; Graefe et al. 2016; Van der Kaa and Krahmer 2014), our participants evaluate the perceived quality of automatic and management summaries. From this literature, an important criterion for evaluating summaries is how well they capture meaningful information in the underlying document being summarized (White et al. 2003). In this preliminary test, MTurk participants compare summaries of two earnings releases to the underlying text of the source documents, judging how well each summary captures the information contained in the earnings release.

For this test, we select two earnings releases—Boeing for Q2 of 2008 and Target for Q4 of 2013—and randomly assign each participant to one of the two earnings releases.[6] Participants first read one of these two earnings releases in full. To facilitate judgments, participants were encouraged to take notes while reading the earnings release, and these notes were reproduced for reference when participants rated the summaries. Each participant then assessed one management summary and four automatic summaries generated by extraction-based summarization algorithms known as KLSum (KL), LexRank (LEX), Latent Semantic Analysis (LSA) and SumBasic (SB) (Nenkova and McKeown 2011).[7] After reading the full earnings release, participants evaluated each summary; we randomized the order in which participants

[6] The Online Appendix compares the tone and readability of these two earnings releases, as well as the earnings release used in the experiment reported in Section 4, to the sample of 78 earnings releases used in our archival analyses. Results suggest that the three earnings releases have net tone (i.e., # of positive words minus # of negative tone words) and Gunning-Fog scores that are reasonably close (within one standard deviation) to the sample mean. Thus, these three earnings releases exhibit characteristics that are representative of earnings releases more generally.

[7] In pretests, reported in the Online Appendix, we tested two additional algorithms, known as Luhn (Luhn 1958) and TextRank (Mihalcea and Tarau 2004), which produced summaries that were significantly longer than either the management summaries or the other automatic summaries. Therefore, we exclude them from our main tests, and focus instead on algorithms that produce summaries that are more similar in length to management summaries.

viewed the summaries.

For each summary, participants provided the following judgments: (1) "Capture"—the extent to which the summary captured important information in the earnings release, (2) "Reliance"—the extent to which participants would rely on the summary in judging the company's performance, (3) "Bias"—the extent to which the summary made the company's performance look better or worse than the underlying earnings release, and (4) "Should be included"—participants' overall preference for whether the summary should or should not be included with the earnings release. "Capture" and "Reliance" judgments were made on 101-point scales with endpoints of 0 and 100 (both endpoints appropriately labeled). "Bias" judgments were made on a 101-point scale with endpoints of –50 ("Summary makes [Company] look worse") and +50 ("Summary makes [Company] look better"). For the "Should be included" judgment, participants selected either "Yes, the summary should be included" or "No, the summary should not be included." Participants also indicated what important information (if any) was missing from each summary, and whether any information included in each summary should not have been included (e.g., because it was redundant or irrelevant).

We recruited participants from Amazon's Mechanical Turk (MTurk) platform (e.g., Rennekamp 2012; Krische 2015; Asay et al. 2017; Farrell et al. 2017; Blankespoor et al. 2017a). The MTurk platform directed potential participants to an online survey designed in Qualtrics. As in other financial accounting studies using MTurk participants (e.g., Bonner et al. 2014), participants were required to pass certain screening questions to ensure that they possess sufficient investment experience to complete the experimental task. Specifically, they were required to be over 18 years of age, to be native English speakers, to have previous investing experience, and to be at least moderately familiar with financial disclosures (indicated by

reported familiarity of 60 or higher out of 100). In total, 334 people volunteered to take part and 98 (29.3%) met the qualification requirements and completed the study. Qualtrics randomly assigned participants to one of the two earnings releases, and the order in which participants viewed the summaries was also randomized. Participants were not told the source of any of the summaries. On completion, participants were paid $4.00 via MTurk. A mean (median) completion time of 39 (24) minutes resulted in a mean (median) hourly rate of $6.15 ($10).[8]

3.1.2 Results

Table 1 reports the results from the user evaluation test. Panel A of Table 1 presents results for the combined sample, and Panels B and C break down results by earnings release. For the combined sample, LexRank appears to be the top performer among the four algorithms, in that it is rated the highest for "Capture," "Reliance," and "Should be included," but is also rated the lowest for "Bias."[9]

< Insert Table 1 about here >

The key finding from this test is that participants judge automatic summaries generated using the LexRank algorithm to be less biased in the company's favor than management summaries. At the same time, participants judge LexRank summaries to capture the important information from earnings releases as well as the management summaries. Further, they are just as likely to rely on LexRank summaries as the management summaries, and to believe that the

[8] Because we used the same procedures to recruit and screen participants for the pretest in the Online Appendix, the user evaluation study reported here, and the experiment reported in Section 4, we report aggregate demographics here. A large majority of our participants (88%) had taken college courses, and 71% held a bachelor's degree or higher. Participants report having taken an average of 3.8 accounting or finance classes, and had an average of 14.5 years of full-time work experience. According to Elliott, Hodge, Kennedy and Pronk (2007), nonprofessional individual investors on average have taken 3.5 accounting or finance courses, and 97 percent have experience with financial statements. Our participants thus had similar profiles to the nonprofessional individual investors in Elliott et al. (2007), suggesting that they were appropriate proxies for nonprofessional individual investors.

[9] LexRank's superior performance is consistent with findings in previous studies (e.g., Verma and Om 2016) and may be attributable to its use of a "reranker." The reranker "penalizes the sentences that are similar to the sentences already included in the summary so that a better information coverage is achieved" Erkan and Radev (2004, 469)

LexRank summaries should be included in the earnings release.[10,11]

Further supporting these conclusions, in untabulated analysis, we compare the number of "fundamental" terms (e.g., "sales", "expenses", "margins") discussed in the body of the earnings release that are also included in the LexRank and management summaries.[12] For Target, of the 16 fundamental terms identified in the underlying text of the earnings release, the LexRank summary includes ten of these terms, compared to seven in the management summary. For Boeing, the LexRank summary includes six of the 12 fundamental terms, compared to four in the management summary.

Overall, these results provide preliminary evidence that it is possible for automatic summaries to capture important information from earnings releases at a level that is comparable to management summaries, but with less bias. With respect to the different summarization algorithms tested, LexRank appears to perform particularly well on these dimensions.

3.2 Intrinsic Evaluation using ROUGE

The above test follows an extrinsic approach, where we rely on user perceptions and use criteria external to the summary to evaluate it (e.g., a summary's perceived bias). Summaries can

[10] Compared to the management summary, for each earnings release, LexRank has lower "Bias" scores and slightly higher "Capture," "Reliance" and "Should be included" scores. Statistically, however, only participants assigned to Target rated "Bias" as significantly lower for the LexRank summary compared to the management summary ($F_{1,56}$ = 12.45, $p < 0.01$). An explanation for this difference is that Target gave too little attention to two major news events—a credit-card breach and a struggling Canadian segment—in the management summary. Three raters (two co-authors and an independent rater) coded how often participants indicated that information related to these events was missing from each summary. In untabulated analyses for the LexRank (management) summary, we find that 19.30% (35.09%) of the participants indicated that important information related to these events was missing. This difference is statistically significant ($F_{1,56}$ = 4.53, p = 0.04) and significantly related to the difference in "Bias" ($F_{1,56}$ = 5.81, p = 0.02). This evidence suggests that Target's management avoided highlighting important negative news events in its summary.

[11] As reported in pretests in the Online Appendix, Table OA.1, LexRank has the highest readability score of all automatic summaries and does not differ compared to management summaries in terms of readability. User evaluation scores further suggest that participants' evaluations of credibility, informativeness, and usefulness are significantly ranked higher for LexRank in comparison to the management summary.

[12] For each earnings release, a list of fundamentals was agreed upon by two of the authors, who independently identified fundamental terms mentioned in the underlying text.

also be evaluated intrinsically (Nenkova and McKeown 2011). Intrinsic evaluation considers the content of the summary relative to an expert reference summary. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is the most commonly used intrinsic evaluation system (Nenkova and McKeown 2011; Moen et al. 2016). Under the ROUGE approach, the distribution of words in the reference summary, Sum_{Ref}, is compared to the distribution of words in candidate summaries, Sum_{Can} (Lin 2004), with a higher ROUGE score indicating a better match between the reference and candidate summaries. Specifically,

$$\text{ROUGE} = \frac{\sum_{\text{Words}} \in \{Sum_{Can} \cap Sum_{Ref}\}}{\sum_{\text{Words}} \in Sum_{Ref}}$$

To construct a reference summary, we rely on an experienced Investor Relations Officer (IRO). The IRO received electronic copies (without summaries) of the same two earnings releases used in the extrinsic user evaluation tests. We asked the IRO to read each earnings release and produce a summary consisting of five sentences (consistent with the number of sentences in the management summary), each presented as a bullet point. We asked that the summary should focus on the important information that best captured the content of the earnings release.

< Insert Table 2 about here >

Table 2 reports ROUGE scores for the combined sample (column 1), Target Q4 2013 (column 2), and Boeing Q2 2008 (column 3), respectively. We find that, in all cases, LexRank summaries have higher ROUGE scores relative to management summaries. That is, the intrinsic analysis reveals that LexRank summaries contain a larger number of words from the set of relevant words identified by the IRO, providing evidence that the LexRank summary better captures elements of the earnings release that are deemed important by the experienced IRO. Further, consistent with the extrinsic user evaluation, LexRank outperforms the other algorithms:

either by being ranked first (for the combined sample, in column 1, or for Target, in column 2) or by being among the group of first-ranked algorithms (for Boeing, in column 3). We note that we find significantly less variation ($p < 0.01$) in ROUGE scores among the four automatic summaries and, on average, significantly higher ROUGE scores ($p < 0.05$) for Boeing compared to Target. The lower variation in ROUGE scores helps explain why we find less variation in extrinsic user evaluations of Boeing summaries in Table 1.

3.3 Archival Analysis: Management's Use of Tone in Earnings Release Summaries

The main finding from the tests performed thus far is that extraction-based summarization algorithms—and LexRank in particular—have the potential to capture important information from earnings releases at a level that is at least comparable to management summaries, but with less bias. In this section, we provide archival evidence to validate this result in a larger sample.

We start by manually collecting all fourth-quarter earnings releases for S&P 100 firms in fiscal 2015. As shown in Panel A of Table 3, a large majority of firms (81%, or 78 of 96 usable earnings releases) include summaries with their earnings releases. Next, for each earnings release that includes a management-generated summary, we generate an automatic summary using LexRank, setting the number of bullet points equal to that of the management summary. To assess language bias, we analyze whether the summaries differ in tone from the underlying text of the earning release, and also whether the management summary and the automatic summary differ in tone. To measure tone, we use the positive and negative tone words from Henry's (2008) context-specific word list, and divide the number of tone words by the total number of words in the summary or the full text (excluding the summary).[13]

[13] In the accounting and finance literature, out of several sentiment word lists, two have been used extensively: Henry (2008), developed in the context of earnings releases, and Loughran and McDonald (2011), created in the context of 10-K filings. (For a discussion of these and other sentiment word lists—e.g., Harvard IV-4, Diction—see Loughran and McDonald 2016.) To empirically motivate our choice of word list, we use regression analysis to

< Insert Table 3 about here >

As shown in Panel B of Table 3, results show that summaries written by management include a smaller percentage of negative words, and a larger percentage of positive words compared to the underlying earnings release (both p-values < 0.01), suggesting that managers engage in tone management in summarizing earnings releases. Conversely, automatic summaries generated using LexRank are less biased in tone. They do not differ significantly in positive tone ($z = 0.50$, $p = 0.62$) and use only slightly fewer negative tone words compared to the underlying text ($z = 1.95$, $p = 0.05$). Perhaps more importantly, when comparing management summaries to automatic summaries, the tone of management summaries is more positive ($z = 4.81$, $p < 0.01$) and less negative ($z = 3.48$, $p < 0.01$) compared to automatic summaries.

Panel C of Table 3 presents multivariate regression analyses, estimated separately for management and automatic summaries, with *Negative tone* (Columns 1-3) and *Positive tone* (columns 4-6) as dependent variables. Independent variables include earnings release, key firm, and corporate governance characteristics. These multivariate results largely confirm the univariate results in Panel B, but further suggest that management and automatic summaries differ primarily in the extent to which they reflect negative, rather than positive, tone in the underlying earnings release.[14]

correlate participants' "Bias" judgments, from the user evaluation test reported earlier, with an earnings release's net sentiment score (positive word frequencies minus negative word frequencies), calculated using either the Henry (2008) or the Loughran and McDonald (2011) sentiment word list. As reported in the Online Appendix, Table OA.2, the coefficients on *Net sentiment* are statistically significantly positive only when using the Henry (2008) word list.

[14] As regards firm and corporate governance characteristics, this analysis yields several interesting insights. For instance, when earnings surprises are larger, the management summary contains fewer negative words compared to the automatic summary. For firms operating in litigious industries, management, on the one hand, uses more cautious language (i.e. fewer negative tone words) relative to the algorithm summary, yet, at the same time uses more positive tone words; suggesting, that the tone in the algorithm-generated summary is more neutral. When the algorithm extracts a positively-toned sentence from the underlying text of the earnings release, this is more likely for larger firms, whose disclosures tend to be more reliable (e.g. DeFond and Jiambalvo 1991). Finally, management summaries tend to be more negatively toned for firms exhibiting higher leverage and having larger boards.

Overall, the archival tone analysis, for this larger sample of S&P 100 firms' earnings releases, corroborates a key finding from the user evaluation tests. That is, automatic summaries generated using the LexRank algorithm have the potential to summarize the text of the underlying earnings release with less bias compared to manager-provided summaries.

4. The Effect of Summaries on Individual Investors' Judgments

Having established that automatic summaries are less positively biased than management summaries, we next test whether automatic and management summaries have differing effects on individual investors' judgments. We begin this section by developing hypotheses regarding the effect of automatic and management summaries on individual investors' judgments.

4.1 Hypothesis Development

If investors are rational and capable of extracting value relevant information from the underlying disclosure to which they have access, then summaries preceding an underlying source document should have little impact on investors' judgments. However, because individual investors are boundedly rational and display limited attention (Hirshleifer and Teoh 2003; Elliott et al. 2015), we expect summaries to affect their judgments, even when they have access to the full underlying source document.

4.1.1 Management Summary versus Automatic Summary

We first consider the impact of a management summary compared to an automatic summary. As shown in the previous section, management summaries tend to be positively biased, depicting a more favorable image of the company compared to both automatic summaries and the underlying text of the earnings release. We posit that this bias is likely to affect individual investors' judgments via two psychological mechanisms: tone effects and primacy effects.

Tone management can affect individual investors' impressions of a target company. Archival research suggests that investors react to opportunistic use of tone (Huang et al. 2013; Davis and Tama-Sweet 2012). Prior experimental research in accounting also suggests that less sophisticated investors are influenced by the opportunistic use of tone in earnings releases, indicating that judgments about the firm's future earnings performance are more favorable when the earnings release is positively written (Tan et al. 2014). Further, as we document from archival data (Table 3), manager-generated summaries are even more favorable in tone than the underlying documents they summarize. In contrast, an automatic summary is more likely to reflect the tone of the underlying document because it is based on sentence extraction.

In addition, research in psychology suggests that decision makers are prone to primacy effects: that is, the order in which information is presented affects subsequent information processing (Asch 1946; Hogarth and Einhorn 1992; Nisbett and Ross 1980). This theory suggests that encoding initial positive information tends to result in more positive global impressions of a target, relative to when participants first encode initial negative information (Sinclair 1988, 25). For example, Sinclair (1988) explicitly manipulates information order, and shows that participants initially receiving positive information (compared to those who initially receive negative information) tend to make more favorable judgments when reviewing the performance of employees. Further, information retrieval seems to be directionally consistent with information order manipulation, with participants retrieving more positive information when seeing positive information first.

Building on this research in psychology, we argue that a summary is a piece of narrative information that is likely to be read first, and hence more likely to be remembered when constructing a problem representation due to primacy effects (Pennington and Hastie 1986). This

is likely to bias the way that participants subsequently acquire and/or interpret the information they review in the underlying earnings release. Thus, even when the full underlying text is available, investors' judgments may differ when management depicts a more favorable picture of the company in their summary compared to an automatic summary.

In sum, whereas the summarization algorithms tested in this paper rely on sentence extraction, management gives information in the earnings release a positive spin by selectively (deemphasizing) emphasizing (negative) positive words and news items when they generate a summary. When we consider this tone management together with the selective content selection and primacy effects discussed above, we predict in H1 that a management summary will have a more favorable effect on individual investors' valuation judgments compared to an automatic summary generated using an extraction-based summarization algorithm.

H1: Individual investors' valuation judgments will be more favorable when a management summary accompanies an earnings release compared to when an automatic summary is provided.

4.1.2 Management Summary versus No Summary

We also consider the effect of a management summary compared to cases where individual investors do not receive a summary preceding the earnings release. In these cases, we also predict that a management summary will affect individual investors' valuation judgments positively, given that management summaries are likely to be positively biased relative to the information disclosed in the underlying document. Thus, H2 is based on the same theory as H1.

H2: Individual investors' valuation judgments will be more favorable when a management summary accompanies an earnings release compared to when no summary is provided.

4.1.3 Automatic Summary versus No Summary

Ex ante, neither theory nor previous research provide a clear directional prediction for how an automatic summary will affect individual investors' judgments compared to when no summary is provided. In the absence of a summary, individual investors must navigate and process a large amount of information, which can be more difficult when no initial guidance is provided (McDonald and Stevenson 1998). A summary arguably offers some initial guidance. Further, research suggests that a query-based summary—which extracts important content from the whole document, rather than simply displaying the first sentences of an article—facilitates information search (Tombros and Sanderson 1998). However, while an automatic summary may guide individual investors in retrieving relevant information, it may not be completely free from bias, given that it relies on sentence extraction from the text of the underlying earnings release, which can itself be biased. Given the uncertainty about the effects of providing participants with an automatic summary, we leave the effect of an automatic summary compared to no summary as an empirical question.

4.2 Research Design and Materials

Our experiment had a 1 × 3 between-subjects design, with summary type (automatic summary, management summary, or no summary) as the manipulated variable. For this study, we created an earnings release for a hypothetical retail company based on Target Corporation's earnings release for the first quarter of 2016. Following prior literature (e.g., Rennnekamp 2012), we disguised the company's identity so that participants' familiarity with a real company would not influence their valuation or other judgments. However, to preserve the external validity of the earnings release and the summaries, we changed only information that would clearly identify Target or other companies named in the earnings release: the firm name and logo, the location of

its headquarters, names and contact details for company employees, and the name of another firm identified in the earnings release. In addition, to be consistent with the time of year in which the study was administered, the earnings release for the hypothetical company reported results from the third quarter of the fiscal year rather than the first quarter. Other than these changes, the earnings release and the management summary used in the study were identical to Target's actual earnings release and summary. The Appendix contains further details of the experimental materials, including the management and automatic summary.

We then generated the automatic summary, with the same number of bullet points (six) as the management summary. Because of its superior performance in our preliminary tests, we used LexRank to generate the automatic summary. In addition, to ensure that the management and automatic summary exhibited similar characteristics as the summaries used in our preliminary user evaluation and archival tests, we compared the automatic and management summary for two potential sources of bias: content management (i.e., managers highlighting certain news items, while withholding others) and tone management (i.e., managing the tone of words in the management summary).

With respect to content management, the company experienced a 5.4% sales decrease during the quarter compared to the same quarter in the previous year. Management does not mention this sales decrease in its summary, while the automatic summary does include a sentence from the underlying earnings release on the sales decrease. This suggests that management avoided mentioning an important negative news item in its summary.[15]

To compare tone management between the management and automatic summary, we measure the frequency of negative and positive words (as a % of total words used) using the

[15] Media reports on Target's Q1 2016 performance also indicate that the sales decrease was interpreted as both significant and negative (e.g., CNBC 2016; Oyedele 2016; Zacks 2016).

context-specific wordlist from Henry (2008). In addition, following Allee and DeAngelis (2015), we compute a measure of linguistic dispersion from the computational linguistics literature—(average) reduced frequency, or (A)RF—to measure the degree to which tone words are evenly distributed throughout the document. A higher RF (closer to 1) indicates that words are more "evenly" distributed throughout the document, while smaller values of RF indicate a "chunkier" distribution. A more even distribution of tone throughout the narrative reflects a portrayal of good or bad news as pervasive, while a less even distribution isolates the news to fewer components of performance.[16]

< Insert Table 4 about here >

The frequency of negative (positive) words in the underlying text of the earnings release—excluding the headline, management summary, and the tables—is a modest one percent (two percent). Both tone dispersion measures are higher for negative words. Consistent with our earlier analyses, Table 4 documents tone frequencies and dispersion scores for the automatic summary similar to those documented for the full earnings release. In contrast, the frequency of positive words (nine percent) is considerably higher in the management summary than in the underlying earnings release, and (A)RF dispersion measures are also relatively high. Combined with content management, discussed above, this analysis indicates that management positively biases information in this summary compared to both the automatic summary and the full earnings release, just as we found in our user evaluation tests and in our archival tests. Thus, we test individual investors' reactions to management and automatic summaries in a representative setting.

[16] Untabulated analysis shows that both summaries have college-level readability scores. Specifically, the Gunning-Fog score, calculated using Python package "textstat" (https://github.com/shivam5992/textstat), is 14.24 for the automatic summary and 16.54 for the management summary, suggesting that, in relative terms, the automatic summary is slightly more readable.

4.3 Participants and Procedures

Participants were recruited from MTurk, using the same procedure as in our user evaluation tests. In total, 308 people volunteered to take part and 90 (29.2%) met the qualification requirements and completed the study. Qualtrics randomly assigned participants to one of the three summary conditions (automatic, management, or no summary). Participants who viewed a summary were not told whether the summary was generated automatically or by management (Clerwall 2014). On completion, participants were paid $2.00 via MTurk. A mean (median) completion time of approximately 16 (10) minutes resulted in a mean (median) hourly rate of $7.49 ($11.63).

Participants first read background information about the hypothetical firm ("Home Square Stores" or "HSQ"), and then provided an initial valuation judgment for the company's common stock on a 101-point scale with endpoints of 0 ("Very low value") and 100 ("Very high value"). Consistent with prior work (e.g., Asay et al. 2017), eliciting an initial valuation judgment allows us to more precisely measure the effect of our manipulation by measuring the difference between valuation judgments before and after observing the earnings release and accompanying summary (or lack of summary). This procedure controls for individual differences, such as participants' views about the retail industry as an investment and their use of the scale.

After making the initial valuation judgment, participants received HSQ's earnings release for the third quarter of 2016. In the automatic and management summary conditions, participants were asked to first read the summary provided before clicking a button that revealed hyperlinks to the sections and tables of the earnings release. Participants in the no summary condition were asked to click the button when they were ready to proceed. Clicking on any of the hyperlinks opened a new window containing the section or table. After reviewing the earnings release

information, but before moving to the next screen, participants made a final valuation judgment for HSQ's common stock on a 101-point scale that was identical to the scale used for the initial valuation judgment. On the next screen, participants made several additional judgments. First, participants indicated how confident they felt when making their final valuation judgment on a 101-point scale with appropriately labeled endpoints. Participants then indicated—via a free response—which factor was most important to their final valuation judgment, and also indicated up to four additional factors that were important. Further, following the approach in Frederickson and Miller (2004), participants rated HSQ's future earnings growth potential, the risk of investing in HSQ's common stock, and the favorability and credibility of HSQ's earnings release, all on 101-point scales with appropriately labeled endpoints.

4.4 Results

Table 5 reports results. Panel A provides descriptive statistics by summary condition. Panel B shows planned comparisons between summary conditions. Based on our hypotheses, we expect higher valuation judgments when participants receive the earnings release with management's summary compared to the automatic summary (H1) or no summary (H2).

< Insert Table 5 about here >

Results support these predictions. With respect to our main dependent measure—the change in participants' valuation judgments (i.e., the final valuation judgment minus the initial valuation judgment)—participants who received the earnings release with management's summary increased their valuation judgments by 7.70 points, compared to an increase of only 0.41 points for those who received the automatic summary ($t_{60} = 1.85$, $p = 0.03$, one-tailed), and a decrease of 0.14 points for those who received no summary ($t_{59} = 2.06$, $p = 0.02$, one-tailed). Valuation judgments of participants who received automatic summaries do not differ

significantly from those of participants who received no summary (t_{55} = 0.13, p = 0.90, two-tailed).[17]

Results for the other investment-related judgments are also reported in Table 5. Consistent with our prediction for the valuation judgment, we expect higher judgments for earnings growth potential and earnings release favorability, and lower risk judgments, when participants receive the earnings release with management's summary compared to the automatic summary or no summary. For the earnings growth potential and earnings release favorability measures, the judgments of participants who received the management summary are at least marginally higher than those of participants who received the automatic summary or no summary (all $p < 0.10$, one-tailed). For the risk measure, participants who received the automatic summary judge the risk of investing in HSQ's common stock to be higher compared to participants who received the management summary, and this difference is marginally significant (t_{60} = 1.35, p = 0.09, one-tailed). However, we observe no difference in the risk judgments of those who received the management summary compared to those who received no summary (t_{59} = 0.03, p = 0.52, one-tailed). The risk judgments of participants who received the automatic summary are directionally, but not significantly, higher than the risk judgments of participants who received no summary (t_{55} = 1.43, p = 0.16, two-tailed).

We also elicited two additional judgments from participants: the credibility of the earnings release, and the confidence they felt when making their final valuation judgments. We observe no differences in the credibility of the earnings release. However, participants who receive the automatic summary are more confident in their (lower) final valuation judgments than participants who receive the management summary (t_{60} = 1.84, p = 0.07, two-tailed).

[17] Results for valuation judgments are inferentially identical if we instead compare final valuation judgments across summary conditions, controlling for initial valuation judgments.

4.5 Mediation Analysis

The analysis reported above indicates that summary type had a significant effect on four of the additional investment-related measures: earnings growth potential, earnings release favorability, risk, and confidence. We next conduct a mediation analysis using Structural Equation Modeling (SEM) to determine which, if any, of these four measures explained investors' valuation judgments.[18] The results of this analysis are presented in Figure 1. Panel A presents results for the effect of the management summary compared to the automatic summary. Panel B presents results for the effect of the management summary compared to no summary.

< Insert Figure 1 about here >

We start by testing the overall goodness of fit for each model. For the management versus automatic model in Panel A, the Tucker-Lewis Index, which measures the improvement in fit compared to a null model, is 1.05, indicating that the model is a good fit for the data. The goodness of fit is confirmed by various other measures, including an Incremental Fit Index of 1.00, and an insignificant χ^2 test ($\chi^2_{(1)} = 0.62$, $p = 0.43$) (Iacobucci 2010; Kline 2011). The management versus no summary model in Panel B is also a good fit for the data, as confirmed by a Tucker-Lewis Index of 1.08, an Incremental Fit Index of 1.00, and an insignificant χ^2 test ($\chi^2_{(1)} = 0.35$, $p = 0.55$).

We next turn to the sign and significance of the path coefficients. Each model includes paths from summary type (the independent variable) to each of the four potential mediators, and paths from each of the four mediators to the change in valuation judgments (the dependent variable).[19] Full mediation is indicated if the following conditions hold: (1) the path from

[18] SEM has advantages over regression in testing mediation, especially in cases that deviate from simple X→M→Y relationships, as is the case in our models with their multiple potential mediators (e.g., Iacobucci et al 2007).
[19] We also allow error terms for the mediators to covary (these covariance paths are omitted from Figure 1 to simplify the presentation).

summary type to the mediator is significant, (2) the path from the mediator to the valuation judgment is significant, and (3) the path from summary type to the change in valuation judgment is insignificantly different from zero with the mediator included in the model (Baron and Kenny 1986; Iacobucci et al. 2007).

Results reveal that, for both models, these conditions are met for two of the potential mediators: earnings growth potential and earnings release favorability. Specifically, the path coefficients for the effect of summary type on earnings growth potential and earnings release favorability are significantly positive. In addition, the path coefficients from these two measures to common stock valuation judgments are significantly positive. Finally, with the potential mediators included in the models, the path coefficients from summary type to common stock valuation are no longer significant. These results indicate that the effect of the management summary on participants' valuation judgments is fully explained by their judgments of earnings growth potential and earnings release favorability.[20]

Two points about these results are in order. First, the mediating effect of potential future earnings growth indicates that participants are sufficiently knowledgeable about the determinants of equity value that their judgments of common stock value are closely related to their judgments of the company's ability to generate future earnings. Second, the mediating effect of earnings release favorability suggests that the management summary changes participants' overall impression of the earnings release, despite receiving the same underlying information, consistent with the theorized primacy and tone effects underlying our hypotheses.

[20] Mediation results are inferentially identical when testing each of the four potential mediators separately, with one exception. Risk judgments partially mediate the effect of summary type (automatic versus management) when risk judgments are tested without the other potential mediators in the model.

4.6 Information Search and Processing

As noted, all participants had access to the full text of the underlying earnings release, and accompanying tables, via hyperlinks. This design allowed us to measure the time participants spent searching specific sections of the earnings release. Further, participants listed up to five factors that were important to their valuation judgment, providing insights into participants' processing of the earnings release information, including the summaries. In this section, we discuss several insights from this information search and processing data.

We first compare, across summary conditions, the time that participants spent searching for and processing information. We measured time in two different ways: total time spent on the study and time spent on the earnings release information. We detect no differences in these time measures across summary conditions (all p-values > 0.10, two-tailed). Next, we compare time spent on each of the earnings release sections and the five tables across summary conditions. Results indicate that participants who received summaries, regardless of summary type, spent significantly more time searching for information in the following sections or tables: "Capital Returned to Shareholders" (section), "Discontinued Operations" (section), "Reconciliation of Non-GAAP Financial Measures" (table), and "Segment Results" (table).

Compared to participants who did not receive a summary, search time for these sections/tables was significantly higher for participants who received automatic summaries and for participants who received management summaries (all p-values < 0.10, two-tailed). Notably, these sections and tables of the earnings release provide detail on the more complex and economically meaningful aspects of the company's performance during the quarter.[21]

[21] For example, the company distributed an unusually large amount of cash—more than $1.2 billion, representing more than twice net income from continuing operations from the quarter—to shareholders during the quarter, either as dividends or as share repurchases. This information was included in the "Capital Returned to Shareholders" section.

Interestingly, we do not observe any significant effects of summary type (i.e., automatic versus management) on search time related to these sections (all *p*-values > 0.10, two tailed). These results suggest that providing a summary, regardless of whether generated automatically or by management, improves the efficiency of participants' information search, directing their focus toward more important or complex sections of the earnings release.

We also coded the influential factors listed by participants for mentions of "sales" or "revenue." As previously noted, the company reported a year-over-year sales decline in the quarter, which was widely interpreted as significant and negative news. The sales decline was explicitly mentioned in the automatic summary, but not in the management summary. Perhaps surprisingly, then, although 35% of participants mention "sales" or "revenue," we do not observe differences across summary conditions ($\chi^2_{(2)} = 1.00$, $p = 0.61$).[22] Combined with the fact that we do observe differences in valuation and other investment-related judgments, this finding suggests that summary type does not seem to affect the acquisition of information from the underlying document, but rather affects the processing and interpretation of the information.

Finally, closer inspection of the data reveals considerable variation in (a) the number of factors listed by participants, and (b) the number of words used to describe factors. Further analysis (untabulated) shows that participants who place higher values on the company's stock are more likely to list a larger number of factors ($p = 0.06$, two-tailed). Building on this finding, for the management and automatic summary condition (N = 61), we regress the net difference of positive and negative tone words (*Net*) on the average number of words used by participants used to describe factors (*Words*), a dummy variable equal to one in case of an automatic summary

[22] We note, however, that these results should be interpreted with caution, as we do not look at qualifiers that accompany these words, and some participants mention sales in an ambiguous way (e.g., without stating explicitly whether sales influenced their judgments positively or negatively).

(*Auto*), and an interaction term *Words* × *Auto*. We find a significantly positive coefficient on *Words* ($p < 0.01$). Interestingly, and consistent with our earlier results, this positive link is attenuated for participants who receive the earnings release accompanied by an automatic summary ($p = 0.10$, two-sided).

5. Conclusion

Automatic summarization technology is today recognized as a useful tool in various disciplines. In this paper, we use multiple methods to investigate how automatic, algorithm-generated summaries compare to management summaries on several dimensions (e.g., bias, reliance), and how summaries affect individual investors' judgments. Our study thus responds to the call by Barth (2015, 506) for research on summaries "to aid investors and other outside providers of capital in their decision making."

Our archival analysis shows that earnings releases often include summaries. However, our tests also reveal that these summaries, provided by managers, introduce incremental positive bias compared to the text of the underlying earnings releases they summarize. Automatic summaries, and those generated by LexRank in particular, have the potential to reduce this bias without sacrificing usefulness. Results also suggest that summaries affect investors' valuation-related judgments. As such, our study informs policy makers, including the SEC, which now explicitly allows filers to provide summary information (SEC 2016). Our results suggest that encouraging management-generated summaries will not necessarily lead to the most value-relevant information being highlighted in a neutral way, and that extraction-based automatic text summarization could be a viable alternative in the context of earnings releases.

This study opens up avenues for future research on the role played by summarization in capital markets. For example, in practice, individual investors could generate an automatic

summary and use it alongside a summary provided by management. When automatic summaries differ from management summaries, managerial bias (i.e., tone and/or content management) might become more evident, which could affect individual investors' interpretation of the underlying information. At the same time, given individual investors' tendency to disregard the content of earnings releases (Blankespoor et al. 2017b), automatic summaries may be better at enhancing "the ability of investors and other users to process relevant information and/or reducing their processing time and search costs" (SEC 2016). In this regard, our supplementary analysis indicates that summaries in general facilitate information search in the relevant sections of an earnings release.

Research could also examine how the existence or widespread use of automatic summaries affects management summaries or indeed the underlying source documents. If managers are aware that a disclosure will be summarized automatically, they might alter their own summary to be less biased and/or change the disclosure itself so that more positive information is identified by the automatic summarization algorithm. Also, as approaches to abstraction advance to better emulate human summarization, future research could investigate the usefulness of abstractive summarization, which requires more sophisticated processing (e.g., semantic interpretation, generation of summary language). Finally, different types of investors have different information needs (e.g., Hales et al. 2011). As automatic summarization technology matures, research can examine automatic summaries that are customized for investors' preferences and/or for other types of disclosures (e.g., conference calls, prospectuses) for which investors may find summaries useful.

APPENDIX
Experimental Materials

Panel A: Earnings Release Header

FOR IMMEDIATE RELEASE

Contacts: James Connelly, Investors, (726) 616-7216
Emily Hubble, Media, (726) 829-5167
Home Square Media Hotline, (726) 696-4300

Home Square Reports Third Quarter 2016 Earnings

Panel B: Management Summary

- *Third quarter Adjusted EPS of $1.29 was above the company's guidance of $1.15 to $1.25.*
- *Third quarter comparable sales increased 1.2 percent, driven by growth in both traffic and basket.*
- *Comparable digital channel sales increased 23 percent, on top of 38 percent growth in third quarter 2015.*
- *Third quarter comparable sales in signature categories (Style, Baby, Kids and Wellness) grew more than three times as fast as the company average.*
- *The third quarter marked Home Square's sixth consecutive quarter of traffic growth, reflecting increases in both stores and digital channels.*
- *Home Square returned $1.2 billion to shareholders in the third quarter through dividends and share repurchases.*

Panel C: Automatic Summary

- *Third quarter GAAP earnings per share (EPS) from continuing operations were $1.02, compared with $1.01 in third quarter 2015.*
- *Third quarter 2016 GAAP EPS from continuing operations reflects $261 million of pre-tax early debt retirement losses, costs related to the sale of the pharmacy and clinic businesses to DA Pharma and the resolution of income tax matters.*
- *Third quarter 2016 sales decreased 5.4 percent to $16.2 billion from $17.1 billion last year, as a 1.2 percent increase in comparable sales was more than offset by the impact of the sale of the pharmacy and clinic businesses.*
- *The Company's third quarter 2016 net interest expense was $415 million, compared with $155 million last year, driven by a $261 million charge related to the early retirement of debt.*
- *Third quarter 2016 effective income tax rate from continuing operations was 31.6 percent, compared with 34.8 percent last year.*
- *Third quarter net earnings from discontinued operations were $18 million, compared with after-tax losses of ($16) million last year.*

Panel D: Earnings Release Sections

Sections

- Preface
- Fiscal 2016 Earnings Guidance
- Segment Results
- Interest Expense and Taxes from Continuing Operations
- Capital Returned to Shareholders
- Discontinued Operations
- Conference Call Details
- Miscellaneous
- About Home Square

Tables

- Consolidated Financial Statements:
  - Consolidated Statements of Operations
  - Consolidated Statements of Financial Position
  - Consolidated Statements of Cash Flows
- Segment Results
- Reconciliation of Non-GAAP Financial Measures

Note: After viewing the earnings release header and summary (header only for no summary condition), participants clicked a button labeled ‘access earnings release sections’, which displayed hyperlinks to the different parts of the company’s full earnings release, as shown in Panel D.

REFERENCES

Ahern, K.R., & Sosyura, D. (2014). Who writes the news? Corporate press releases during merger negotiations. *The Journal of Finance*, 69, 241-291.

Allee, K., & DeAngelis, M. (2015). The structure of voluntary disclosure narratives: Evidence from tone dispersion. *Journal of Accounting Research*, 53, 241-274.

Arnold, V., Bedard, J.C., Phillips, J.R., & Sutton, S.G. (2014). The impact of tagging qualitative financial information on investor decision making: Implications for XBRL. *International Journal of Accounting Information Systems*, 13, 2-20.

Arslan-Ayaydin, O., Boudt, K., & Thewissen, J. (2016). Managers set the tone: Equity incentives and the tone of earnings press releases. *Journal of Banking & Finance*, 72, S132-S147.

Asay, H.S., Elliott, W.B., & Rennekamp, K.R. (2017). Disclosure readability and the sensitivity of investors' valuation judgments to outside information. *The Accounting Review,* 92, 1-25.

Asch, S.E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, 41, 258-290.

Baron, R.M., & Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182.

Barth, M.E. (2015). Financial accounting research, practice, and financial accountability. *ABACUS: A Journal of Accounting, Finance and Business Studies*, 51, 499-510.

Blankespoor, E., Hendricks, B.E., & Miller, G.S. (2017a). Perceptions and price: Evidence from CEO presentations at IPO roadshows. *Journal of Accounting Research*, 55, 275-327.

Blankespoor, E., deHaan, E., Wertz, J., & Zhu, C. (2017b). *Why do some investors disregard earnings news? Awareness costs versus processing costs*. Working paper.

Blankespoor, E., deHaan, E., & Zhu, C. (2018). Capital market effects of media synthesis and dissemination: Evidence from robo-journalism. *Review of Accounting Studies*, 23, 1-36.

Bonner, S.E., Clor-Proell, S.M., & Koonce, L. (2014). Mental accounting and disaggregation based on the sign and relative magnitude of income statement items. *The Accounting Review*, 89, 2087-2114.

Brandow, R., Mitze, K., & Rau, L.E. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing & Management*, 31, 675-685.

Bushee, B.J., Gow, I.D., & Taylor, D.J. (2017). Linguistic complexity in firm disclosures: Obfuscation or information? *Journal of Accounting Research*, forthcoming.

Clerwall, C. (2014). Enter the robot journalist. *Journalism Practice*, 8, 519-531.

CNBC. (2016). *Target earnings top expectations, but revenue is light*. Retrieved from http://www.cnbc.com/2016/05/18/target-q1-earnings-report.html.

Daniel, K., Hirshleifer, D., & Teoh, S.H. (2002). Investor psychology in capital markets: Evidence and policy implications. *Journal of Monetary Economics*, 49, 139-209.

Davis, A., & Tama-Sweet, I. (2012). Managers use of language across alternative disclosure outlets: Earnings press releases versus MD&A. *Contemporary Accounting Research*, 29, 804–837.

DeFond, M., & Jiambalvo, J. (1991). Incidence and circumstances of accounting errors. *The Accounting Review*, 66, 643-655.

DeFranco, G., Fogel-Yaari, H., & Li, H. (2016). *MD&A textual similarity and auditors*. Working paper.

Desai, H., Rajgopal, S., & Yu, J.J. (2016). Were information intermediaries sensitive to the financial statement-based leading indicators of bank distress prior to the financial crisis. *Contemporary Accounting Research*, 33, 576-606.

Dyck, A., & Zingales, L. (2003). *The media and asset prices*. Working paper.

Dyer, T., Lang, M., and Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics*, 64, 221-245.

Elliott, W.B., Hobson, J.L., & White, B.J. (2015). Earnings metrics, information processing, and price efficiency in laboratory markets. *Journal of Accounting Research*, 55, 555-592.

Elliott, W.B., Hodge, F.D., Kennedy, J.J., & Pronk, M. (2007). Are MBA students a good proxy for nonprofessional investors? *The Accounting Review*, 82, 139-168.

Erkan, G., & Radev, D.R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22, 457-479.

Farrell, A., Grenier, J.H., & Leiby, J. (2017). Scoundrels or stars? Theory and evidence on the quality of workers in online labor markets. *The Accounting Review*, 92, 93-114.

FASB. (2015). *FASB's simplification initiative: An update*. Retrieved from http://www.fasb.org/jsp/FASB/Page/SectionPage&cid=1176165963019

Frederickson, J., & Miller, J. (2004). The effects of pro forma earnings disclosures on analysts' and nonprofessional investors' valuation judgments. *The Accounting Review*, 79, 667-686.

Francis, J., Schipper, K., & Vincent, L. (2002). Expanded disclosures and the increased usefulness of earnings announcements. *The Accounting Review*, 77, 515-546.

Ganesan, K. (2015). *ROUGE 2.0: Updated and improved measures for evaluation of summarization tasks*. Retrieved from https://arxiv.org/abs/1803.01937.

Graefe, A., Haim, M., Haarmann, B., & Brosius, H-B. (2016). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism*, 1-16.

Guillarmon-Saorin, E., Garcia Osma, B., & Jones, M.J. (2012). Opportunistic disclosure in press release headlines. *Accounting and Business Research*, 42, 143-168.

Hales, J., Kuang, X., & Venkataraman, S. (2011). Who believes the hype? An experimental investigation of how language affects investor judgments. *Journal of Accounting Research*, 49, 223-255.

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *Journal of Business Communications*, 45, 363–407.

Henry, E., & Leone, A. (2016). Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. *The Accounting Review*, 91, 153-178.

Hirshleifer, D., & Teoh, S.H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics*, 36, 337-386.

Hirshleifer, D., Lim, S.S., & Teoh, S.H. (2009). Driven to distraction: Extraneous events and underreaction to earnings news. *The Journal of Finance*, 64, 2289-2325.

Hogarth, R.M., & Einhorn, H.J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24, l-55.

Huang, X., Nekrasov, A., & Teoh, S.H. (2013). *Headline salience and over- and underreactions to earnings*. Working paper.

Huang, X., Teoh, S.H., & Zhang, Y. (2014). Tone management. *The Accounting Review*, 89, 1083-1113.

Iacobucci, D., Saldanha, N., & Deng, X. (2007). A meditation on mediation: Evidence that structural equations models perform better than regression. *Journal of Consumer Psychology*, 17, 139-153.

Iacobucci, D. (2010). Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology*, 20, 90-98.

Kline, R.B. (2011). *Principles and practice of structural equation modeling*. 3rd edition. New York: The Guilford Press.

KPMG. (2011). *Disclosure overload and complexity: Hidden in plain sight*. Retrieved from https://www.kpmg.com/US/en/IssuesAndInsights/ArticlesPublications/Documents/disclosure-overload-complexity.pdf.

Krische, S.D. (2015). *The impact of individual investors' financial literacy on assessments of conflicts of interest.* Working paper.

Kuvaas, B., & Selart, M. (2004). Effects of attribute framing on cognitive processing and evaluation. *Organizational Behavior and Human Decision Processes*, 95, 198-207.

Levin, I.P., Schneider, S.L., & Gaeth, G.J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, 76, 149-188.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66, 35-65.

Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance*, 69, 1643-1671.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54, 1187-1230.

Luhn, H.P. (1958). The automatic creation of literature abstracts, *IBM Journal*, April, 159-165.

MacGregor, D. G., Slovic, P., Dreman, D., & Berry, M. (2000). Imagery, Affect, and Financial Judgment. *Journal of Psychology and Financial Markets*, 1, 104-110.

McDonald, S., & Stevenson, R.J. (1998). Navigation in hyperspace: An evaluation of the effects of navigational tools and subject matter expertise on browsing and information retrieval in hypertext. *Interacting with Computers*, 10, 129-142.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. UNT Digital Library. Retrieved from http://digital.library.unt.edu/ark:/67531/metadc30962/. Accessed December 19, 2016.

Moen H., Peltonen, L-M., Heimonen, J., Airola, A., Pahikkala, T., Salakoski, T., & Salanterä, S. (2016). Comparison of automatic summarization methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67, 25–37.

Nenkova, A. & McKeown, K. (2011). Automatic summarization. *Foundations and Trends in Information Retrieval*, 5, 103-233.

Newman, M.E.J. (2008). The mathematics of networks. *The New Palgrave Dictionary of Economics*, 2nd edition, edited by S.N. Durlauf and L.E. Blume.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice Hall.

Oyedele, A. (2016). Target says its sales will stink. *Business Insider*. Retrieved from http://www.businessinsider.com/target-earnings-2016-5.

Paredes, T. (2003). Blinded by the light: Information overload and its consequences for securities regulation. *Washington University Law Quarterly*, 81, 417-485.

Paredes, T. (2013). Remarks at the SEC speaks in 2013. Retrieved from www.sec.gov/News/Speech/Detail/Speech/1365171492408.

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology*, 51, 242-258.

Rennekamp, K.R. (2012). Processing fluency and investors' reactions to disclosure readability. *Journal of Accounting Research*, 50, 1319-1354.

Salton, G., Sighhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing & Management*, 33, 193-207.

SEC. (2013). *Report on review of disclosure requirements in Regulation S-K*. Retrieved from https://www.sec.gov/news/studies/2013/reg-sk-disclosure-requirements-review.pdf.

SEC. (2016). *Release no. 34-77969: Form 10-K summary*. Retrieved from https://www.sec.gov/rules/interim/2016/34-77969.pdf.

Sinclair, R.C. (1988). Mood, categorization breadth, and performance appraisal: The effects of order of information acquisition and affective state on halo, accuracy, information retrieval, and evaluations. *Organizational Behavior and Human Decision Processes*, 42, 22-46.

Tan, H-T., Wang, E.Y., & Zhou, B. (2014). When the use of positive language backfires: The joint effect of tone, readability, and investor sophistication on earnings judgments. *Journal of Accounting Research*, 52, 273-302.

Tombros, A. & Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval,* 2-10.

Umar, T. (2017). Complexity aversion when Seeking Alpha. *Working paper*.

Van der Kaa, H., & Krahmer, E. (2014). Journalist versus news consumer: The perceived credibility of machine written news. In: *Proceedings of the Computation + Journalism Conference*. New York.

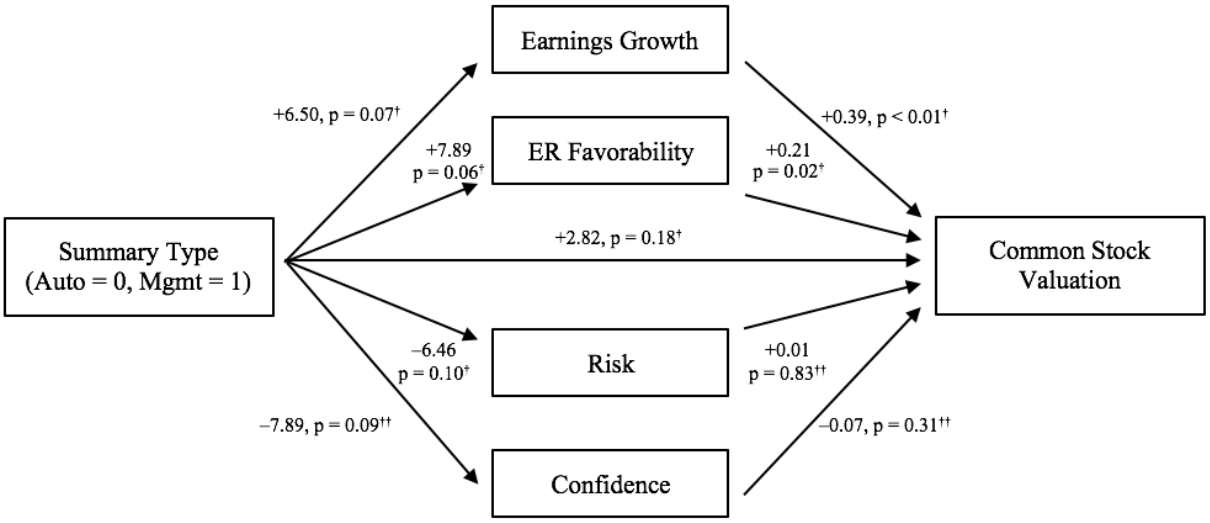
Verma, P., & Om, H. (2016). Extraction based text summarization methods on user's review data: A comparative study. In: Unal A., et al. (eds) "Smart trends in information technology and computer communications." *Communications in Computer and Information Science*, 628, 346-354.

White, R.W., Jose, J.M., & Ruthven, I. (2013). A task-oriented study on the influencing effects of query-biased summarization in web searching. *Information Processing and Management*, 39, 707–733.

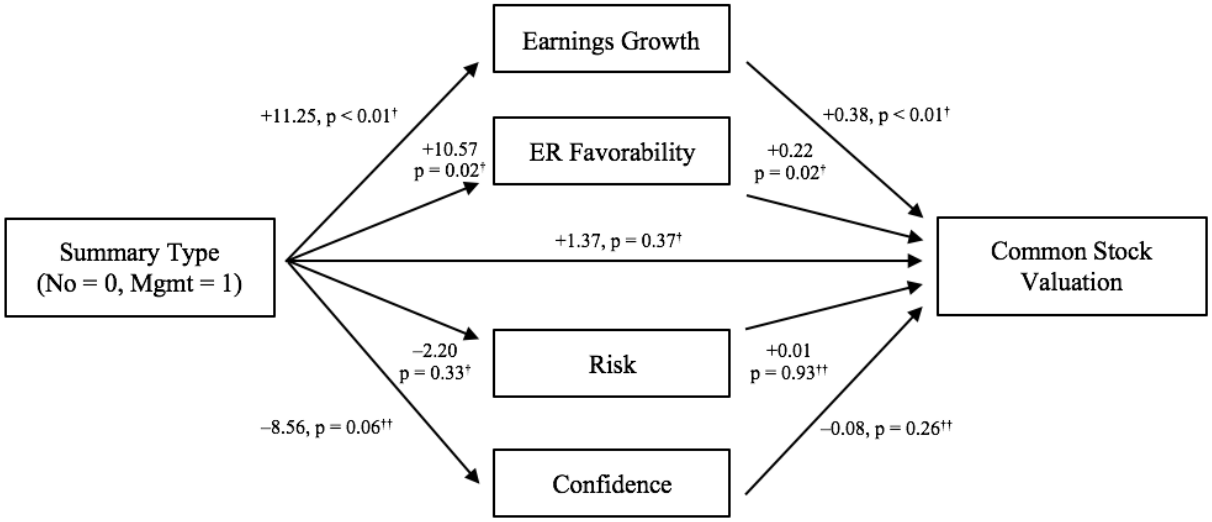
Zacks. (2016). Target (TGT) Q1 earnings beat, sales miss; Stock plunges. Retrieved from https://www.zacks.com/stock/news/217811/target-tgt-q1-earnings-beat-sales-miss-stock-plunges.

FIGURE 1
Experiment: Mediation Analysis

Panel A: Management versus Automatic Summary



Panel B: Management versus No Summary



Panel A (B) presents results of a structural equation analysis that tests potential mediators of the effect of a management summary compared to an automatic summary (no summary) on participants' judgments of common stock value. Next to each arrow are path coefficients and *p*-values (with † and †† indicating one-tailed and two-tailed tests, respectively). Overall goodness of fit is high for both models, as measured by the following measures. Panel A: Tucker-Lewis Index (1.05), Incremental Fit Index (1.00), χ^2 test ($\chi^2_{(1)} = 0.62$, $p = 0.43$). Panel B: Tucker-Lewis Index (1.08), Incremental Fit Index (1.00), χ^2 test ($\chi^2_{(1)} = 0.35$, $p = 0.55$).

TABLE 1
User Evaluation of Automatic and Management Summaries

Panel A: Combined, N = 98

	(1) KL	(2) LEX	(3) LSA	(4) SB	Average Automatic	(5) Management
Capture	57.70	64.39	56.77*	52.77***	57.91	61.70
Reliance	55.39	61.27	55.00	50.50***	55.54	59.11
Should be included (% yes)	65.3%	78.6%	61.2%*	56.1%**	65.3%	73.5%
Bias	**6.08**	**4.01*****	**8.83**	**5.81***	**6.18****	**11.14**

Panel B: Target Q4 2013, N = 57

	(1) KL	(2) LEX	(3) LSA	(4) SB	Average Automatic	(5) Management
Capture	55.28**	64.88	53.86***	47.84***	55.46***	64.05
Reliance	53.04**	63.65	52.61***	46.51***	53.95***	61.19
Should be included (% yes)	59.6%	77.2%	56.1%**	49.1%**	60.6%	73.7%
Bias	–0.53***	3.19***	**14.25**	0.09***	**4.25*****	**13.09**

Panel C: Boeing Q2 2008, N = 41

	(1) KL	(2) LEX	(3) LSA	(4) SB	Average Automatic	(5) Management
Capture	61.07	63.71	60.80	59.61	61.30	58.44
Reliance	58.66	58.00	58.32	56.05	57.76	56.22
Should be included (% yes)	73.2%	80.5%	68.3%	65.9%	72.0%	73.2%
Bias	**15.27**	**5.15**	1.29	**13.76**	**8.87**	**8.44**

"Capture" and "Reliance" judgments were made on 101-point scales with endpoints of 0 and 100 (both endpoints appropriately labeled). "Bias" judgments were made on a 101-point scale with endpoints of –50 ("Summary makes [Company] look worse") and +50 ("Summary makes [Company] look better"). For the "Should be included" judgment, participants selected either "Yes, the summary should be included" or "No, the summary should not be included." "Bias" is reported in **bold** if it is statistically significantly different from zero at the 10% level (two-tailed) or better. *,**,*** indicate statistically significantly different from management summary at $p < 0.10$, $p < 0.05$ and $p < 0.01$, respectively (all two-tailed). Shaded cells indicate where LexRank, compared to the other summarization algorithms, obtained the highest user evaluation on "Capture," "Reliance," and "Should be included," and the lowest user evaluation on "Bias."

TABLE 2
Intrinsic Evaluation: ROUGE Analysis

	(1) Combined	(2) Target Q4 2013		(3) Boeing Q2 2008	
KL	0.3948	0.3076	average	**0.4821**	average
LEX	**0.4590**	**0.4359**	0.3364	**0.4821**	0.4776**
LSA	0.3884	0.2948	st.dev.	**0.4821**	st.dev.
SB	0.3859	0.3076	0.0665	0.4642	0.0089***
Management	0.4129	0.3974		0.4285	

ROUGE scores are obtained using ROUGE 2.0, a Java package developed by Kavita Ganesan (https://github.com/RxNLP/ROUGE-2.0/), where we allow for synonyms and remove stop words during ROUGE scoring (Ganesan 2015). For comparison purposes, we generate an expert summary as reference summary (Moen et al. 2016). To this end, we rely on an experienced Investor Relations Officer, that was instructed to produce a summary that best captured the content of the earnings release using five sentences, that each should be presented as a bullet point. Automatic summaries were generated using KLSum (KL), LexRank (LEX), Latent Semantic Analysis (LSA), and SumBasic (SB). The management summary was taken from the respective earnings release. To identify synonyms, we used the latest version of WordNet (wordnet.princeton.edu/wordnet/download/current-version). Per column, the summary with the highest recall is reported in **bold**. *,**,*** indicate that the difference in average (standard deviation in) ROUGE scores between the two cases — Target Q4 2013 and Boeing Q2 2008 — is statistically significantly at $p < 0.10$, $p < 0.05$ and $p < 0.01$, respectively (all two-tailed).

TABLE 3
Management's Use of Tone in S&P 100 Firms' Earnings Release Summaries

Panel A: Sample Selection

	Q4-2015 earnings release
All S&P 100 firms	100
Document format unsuitable for analysis	(4)
Subtotal: disclosures available for analysis	96
Disclosures not including summaries	(18)
Disclosures including summaries available for analysis	78
Full text, excluding summary: median word count	2,702
Summary: median word count	107

Panel B: Underlying Text, Management Summary, and Automatic Summary (N = 78)

	Underlying text	Management summary	Automatic summary	Difference = (2) – (1)	Difference = (3) – (1)	Difference = (3) – (2)
	(1)	(2)	(3)	(4)	(5)	(6)
Negative tone	0.82%	0.00%	0.56%	-0.82%***	-0.26%*	0.56%***
				(-6.11)	(-1.95)	(3.48)
Positive tone	1.91%	3.83%	2.20%	1.92%***	0.29%	-1.63%***
				(5.82)	(0.50)	(-4.81)

TABLE 3
Management's Use of Tone in S&P 100 Firms' Earnings Release Summaries — *Continued*

Panel C: Cross-Sectional Determinants (N = 78)

	Negative tone (summary)			Positive tone (summary)		
	Management summary	Automatic summary	Equal coeff. (1) = (2)	Management summary	Automatic summary	Equal coeff. (4) = (5)
	(1)	(2)	(3)	(4)	(5)	(6)
Negative tone (earnings release)	0.5616***	1.2409***	4.89**			
	(0.21)	(0.25)				
Positive tone (earnings release)				1.3740***	1.1164***	0.33
				(0.47)	(0.18)	
Readability (earnings release)	0.0022	0.0257	0.20	0.0287	-0.0054	0.07
	(0.03)	(0.04)		(0.14)	(0.05)	
Firm size	0.0253	-0.0850	0.36	-0.3081	0.3371**	4.67**
	(0.17)	(0.13)		(0.33)	(0.14)	
High market-to-book (0,1)	-0.1005	-0.8918*	1.21	0.7245	0.5234	0.02
	(0.74)	(0.50)		(2.13)	(0.62)	
Loss making (0,1)	-0.2334	0.0533	0.49	-0.5107	-0.6921	0.02
	(0.29)	(0.34)		(1.17)	(0.46)	
Litigious industry (0,1)	0.6660**	-0.4105**	11.85***	0.8836	-0.6001	3.40*
	(0.28)	(0.19)		(0.74)	(0.36)	
SUE	-7.5483	0.6597	7.67***	-2.3474	-2.3763	0.00
	(6.73)	(5.64)		(17.7)	(10.0)	
Leverage	1.5874*	-0.6276	3.95**	-0.1389	-0.7506	0.05
	(0.89)	(0.80)		(2.81)	(0.94)	
Independent directors (%)	-0.0761	-1.5585	0.32	-5.0138	-1.7925	0.25
	(1.98)	(1.81)		(7.82)	(1.97)	
Board size	0.1462***	-0.0456	8.72***	0.1164	-0.1559***	4.87**
	(0.05)	(0.03)		(0.13)	(0.05)	
Powerful CEO (0,1)	-0.3410	0.0747	1.15	-0.1696	1.8566	0.71
	(0.35)	(0.48)		(2.41)	(1.63)	
Intercept	-3.1312	3.5100		6.8624	-0.0347	
	(4.20)	(2.73)		(9.98)	(2.80)	
Wald Chi-square	31.64	82.27		19.90	69.78	
	($p < 0.01$)	($p < 0.01$)		($p < 0.05$)	($p < 0.01$)	
R-square	31.23%	53.11%		13.02%	51.82%	

TABLE 3
Management's Use of Tone in S&P 100 Firms' Earnings Release Summaries — *Continued*

For the 78 earnings releases that contain a management summary, we generate an automatic summary using the LexRank algorithm, setting the number of sentences equal to the number of sentences in the management summary for each earnings release. Based on the analysis reported in Table OA.2 in the Online Appendix, we measure tone using the positive and negative tone words from Henry's (2008) context-specific word list. In Panel B, the percentages are expressed against the total number of words in either the underlying text of the earnings release, management summary or automatic summary. ***, **, * indicate relevant Wilcoxon rank-sum (Mann-Whitney) tests (test statistics in parentheses) of differences at $p < 0.10$, $p < 0.05$ and $p < 0.01$, respectively (all two-tailed). Panel C reports coefficient estimates from ordinary least squares regressions, with standard errors (reported in parentheses) bootstrapped using 100 random draws. *Negative tone* (*Positive tone*) is the number of negative (positive) tone words scaled by total words in the underlying text of the earnings release (ER), management summary, or automatic summary, respectively. The dependent variables and the independent variables, *Negative tone (earnings release)* and *Positive tone (earnings release)*, are multiplied by 100. *Readability (earnings release)* is the Gunning-Fog index of the underlying text of the earnings release, measured using the Python package textstat. We use the natural logarithm of book value of total assets (in millions of dollars) as a measure of firm size. *High market-to-book* equals 1 if a firm's market-to-book ratio > 1, 0 otherwise. *Loss making* equals 1 if income before extraordinary items (ibq) < 0, 0 otherwise. *Litigious industry* equals 1 if member in a high-risk industry (SICs 2833-2836, 3570-3577, 7370-7374, 3600-3674, 5200-5961), 0 otherwise. *SUE* is measured as ($ibq_{q=4,t} - ibq_{q=4,t-1}$) / market value of equity. *Leverage* is computed as total long-term debt (ltq) divided by total assets (atq). *Independent directors* is count of independent directors in BoardEx (if boardrole = "Independent Director," "Independent NED," "Independent Board Member," "Independent Outside Director") divided by board size. *Board size* is the count of directors in BoardEx. *Powerful CEO* equals 1 if CEO and chairman, 0 otherwise. *SUE* and *Leverage* are winsorized at 1st and 99th percentile due to concerns with outliers. Columns (3) and (6) report Chi-square statistics from a Wald test of coefficient equality. ***, **, * indicate statistical significance at the 1, 5, and 10% levels, respectively.

TABLE 4
Experiment: Linguistic Tone

	Earnings release full text	Management summary	Automatic summary
	(1)	(2)	(3)
Panel A: Negative Tone Words			
Frequency	5	0	1
(as % of total words)	(1%)	(0%)	(1%)
Negative RF	0.60	N/A	1.00
Negative ARF	0.74	N/A	1.00
Panel B: Positive Tone Words			
Frequency	23	9	2
(as % of total words)	(2%)	(9%)	(1%)
Positive RF	0.52	0.78	0.50
Positive ARF	0.54	0.69	0.53

The earnings release is based on Target Corporation's earnings release for the first quarter of 2016. Motivated by the analysis reported in Table OA.2 in the Online Appendix, we identify positive and negative tone words using the context-specific wordlist developed by Henry (2008). Following Allee and DeAngelis (2015), (average) reduced frequency, or (A)RF, measures the degree to which tone words are evenly distributed throughout the document. A higher RF (closer to 1) indicates that words are more "evenly" distributed throughout the document, while smaller values of RF indicate a "chunkier" distribution.

TABLE 5
Experiment: Hypothesis Testing

Panel A: Mean (Standard Deviation) of Participants' Judgments

	Summary type		
Judgment	Automatic N = 29	Management N = 33	None N = 28
	(1)	(2)	(3)
Initial valuation (pre-manipulation)	55.72	53.61	54.46
	(12.81)	(14.73)	(10.36)
Final valuation (post-manipulation)	56.14	61.30	54.32
	(18.59)	(18.49)	(18.77)
Change in valuation (final minus initial)	0.41	7.70	–0.14
	(16.87)	(14.09)	(15.67)
Earnings growth potential	58.21	64.45	53.71
	(18.11)	(18.51)	(17.18)
Earnings release favorability	56.90	64.67	54.89
	(24.90)	(18.79)	(18.21)
Risk	59.55	52.21	52.04
	(21.29)	(21.50)	(18.33)
Earnings release credibility	72.76	68.94	71.96
	(15.31)	(16.64)	(14.89)
Confidence	73.31	63.45	69.89
	(20.88)	(21.27)	(12.65)

Panel B: Comparisons

Judgment	Contrast	Expectation	*t*-stat	*p*-value
Change in valuation	Mgmt vs. Auto	Mgmt > Auto	1.85	0.03†
	Mgmt vs. None	Mgmt > None	2.06	0.02†
	Auto vs. None	?	0.13	0.90††
Earnings growth potential	Mgmt vs. Auto	Mgmt > Auto	1.34	0.09†
	Mgmt vs. None	Mgmt > None	2.33	0.01†
	Auto vs. None	?	0.96	0.34††
Earnings release favorability	Mgmt vs. Auto	Mgmt > Auto	1.40	0.09†
	Mgmt vs. None	Mgmt > None	2.05	0.02†
	Auto vs. None	?	0.35	0.73††
Risk	Mgmt vs. Auto	Mgmt < Auto	1.35	0.09†
	Mgmt vs. None	Mgmt < None	0.03	0.52†
	Auto vs. None	?	1.43	0.16††
Earnings release credibility	Mgmt vs. Auto	?	0.94	0.35††
	Mgmt vs. None	?	0.74	0.46††
	Auto vs. None	?	0.20	0.84††
Confidence	Mgmt vs. Auto	?	1.84	0.07††
	Mgmt vs. None	?	1.40	0.17††
	Auto vs. None	?	0.74	0.46††

†, †† designate one-tailed and two-tailed *p*-values, respectively.

Online Appendix To
"Automatic summaries of earnings releases: Attributes and effects on investors' judgments"

There are two sections in this Online Appendix. Section 1 contains a primer on automatic summarization. This primer is not intended to be exhaustive. For further discussion, we refer the reader to the original papers referenced herein, and textbooks such as Juan-Manuel Torres-Moreno's *Automatic Text Summarization*, Wiley; Inderjeet Mani's *Automatic Summarization*, John Benjamins Publishing Company; or Inderjeet Mani and Mark Maybury's *Advances in Automatic Text Summarization*, MIT Press. For a review of the literature, see Nenkova and McKeown (2011). Section 2 presents sample descriptives and reports on and tabulates several preliminary tests.

SECTION 1

Text Summarization

A summary is "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that." (Radev et al. 2002). To summarize a text implies taking "an information source, extract[ing] content from it, and present[ing] the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs" (Mani 2001).

Types of Text Summarization

Summarization techniques can be classified into two types: summarization by abstraction and summarization by extraction.

Summarization by Abstraction

Based on semantic understanding, abstraction-based summaries convey the main information in the input, may reuse phrases or clauses from it, expressed in the words of the summarizer (Nenkova and McKeown 2011). In contrast to extraction-based summarization there has been limited research on summarization by abstraction, probably because abstraction-based summarization is beyond the capability of even state-of-the-art automatic summarization techniques. "Very few abstract summarization systems have been created (…). We are (…) a long way off achieving genuine automatic text understanding" (Torres-Moreno 2014, 35).

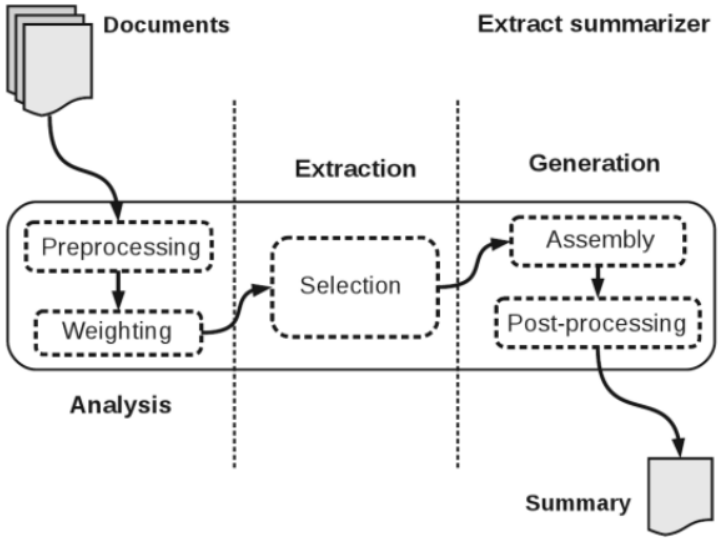
Summarization by Extraction

The essence of extraction-based summarization is to select lexical units containing a document's essential information (i.e., informative content), concatenated into an extractive summary, aiming to give an overview of the original text's content. "Currently, extraction

algorithms dominate the landscape and are at the center of countless automatic summarization systems. The ease with which these methods can be implemented and their good performance are the key to their success" (Torres-Moreno 2014, 271).

Figure OA.1 summarizes the summarization-by-extraction process (figure taken from Torres-Moreno 2014).

FIGURE OA.1
Summarization-by-Extraction Process



The basic idea is to first split a document into lexical units (i.e., sentences). After weighting those using statistical heuristics, the algorithm extracts the units with the highest scores, and assembles them to create a summary.

Extraction-Based Summarization Algorithms

Below we provide some detail on the six algorithms we used to generate the automatic summaries discussed in the paper. In each case, given a text, the summarization task consists in extracting sentences to be included in the summary such that they cover important information with minimal redundancy, while satisfying a length constraint. The algorithms differ in the statistical heuristics (see Figure OA.1 above) applied.

Luhn

The Luhn algorithm—named after its creator, H.P. Luhn: IBM Research Center—collects the frequencies of words in the text and identifies a subset of significant words, excluding the most frequent and the least frequent. The algorithm, then, treats all significant words as having equal weight and computes the weight of a sentence as a function of the concentration of significant words in the sentence (Luhn 1958).

SumBasic

In contrast to Luhn, the SumBasic algorithm relies only on word probability to calculate salience; it uses true initial probabilities and computes the weight of a sentence as equal to the average probability of the words in a sentence (Vanderwende et al. 2007). Specifically, for each sentence S_j in the input, the algorithm assigns a weight equal to the average probability $p(w_i)$ of the content words in the sentence, estimated from the input for summarization:

$$\text{Weight}(S_j) = \frac{\sum_{w_i \in S_j} p(w_i)}{|\{w_i | w_i \in S_j\}|}$$

Then SumBasic picks the best scoring sentence that contains the word that currently has the highest probability. This selection strategy assumes that at each point when a sentence is selected, a single word—that with highest probability—represents the most important topic in the document and the goal is to select the best sentence that covers this word. After the best sentence is selected, the probability of each word that appears in the chosen sentence is adjusted. It is set to a smaller value, equal to the square of the probability of the word at the beginning of the current selection step, to reflect the fact that the probability of a word occurring twice in a summary is lower than the probability of the word occurring only once. This selection loop is repeated until the desired summary length is achieved.

TextRank, LexRank

In graph-based summarization research, TextRank (Mihalcea and Tarau 2004) and LexRank (Erkan and Radev 2004) are the most well-known and often cited. These methods model text as a graph with sentences as nodes and edges based on word overlap. A sentence node is then ranked according to its similarity with other nodes. Specifically, if a sentence S_i is represented as a set of words:

$$S_i = w_1^i, w_2^i, ...w_{|S_i|}^i$$

then the similarity between two sentences S_i and S_j is defined as:

$$Sim(S_i, S_j) = \frac{|\{w_k : w_k \in S_i \wedge w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)}$$

An edge based on similarity can be seen as a process of "recommendation": a sentence that addresses certain concepts, gives the reader a "recommendation" to refer to other sentences that address the same concepts. The underlying assumption for calculating relevance is that the sentences which are similar to a large number of other important sentences are "central." Finally, PageRank (Brin and Page 1998) is used to calculate a relevance score for each sentence based on the relevance score of its similar sentences. Top ranked sentences are selected for the summary such that their total length satisfies the summary length constraint.

Latent Semantic Analysis

At the heart of the Latent Semantic Analysis (LSA) approach is the representation of the input documents as a word by sentence matrix *A*: each row corresponds to a word that appears in the input and each column corresponds to a sentence in the input. Each entry a_{ij} of the matrix corresponds to the weight of word *i* in sentence *j*. If the sentence does not contain the word, the weight is zero, otherwise the weight is equal to the tf*idf weight of the word. Standard

techniques for singular value decomposition (SVD) from linear algebra are applied to the matrix *A*, to represent it as the product of three matrices:

$$A = U\Sigma V^{\mathrm{T}}$$

The rows of V^T can be regarded as mutually independent topics discussed in the input, while each column represents a sentence from the document. In order to produce an extractive summary, the algorithm consecutively considers each row of V^T, and selects the sentence with the highest value, until the desired summary length is reached (Gong and Liu 2001; Steinberger and Jezek 2004).

KLSum

The KLSum algorithm selects a set of sentences from the source document, *D*, such that the distribution of words in the selected sentences—i.e., the "candidate summary," *S*—is as close as possible to distribution of words in document *D*. Specifically, the algorithm introduces the following selection criterion:

$$\mathbf{S}^* = \min_{\mathbf{S}:words(\mathbf{S})\leq L} KL(P_{\mathcal{D}} \| P_{\mathbf{S}})$$

where P_S (P_D) is the word (i.e., unigram) distribution of candidate summary *S* (document *D*). To measure similarity across the word distributions, P_S and P_D, the Kullback-Lieber (KL) divergence measure is used (Haghighi and Vanderwende 2009).

SECTION 2

Preliminary test

In this preliminary test, we compare users' perceptions of management and automatic summaries for three earnings releases from actual companies. Consistent with the literature in communications research, in which users evaluate algorithm- and human-generated news content (e.g., Van der Kaa and Krahmer 2014), our participants evaluated the perceived quality of automatic and management summaries. Specifically, participants judged summaries on attributes including informativeness, readability, credibility and overall usefulness.

We manipulated summary type by having each participant assess one management summary and six automatic summaries generated by extraction-based summarization algorithms: KLSum (KL), LexRank (LEX), Latent Semantic Analysis (LSA), Luhn (LUHN), SumBasic (SB), and TextRank (TR). We randomized the order in which participants viewed the summaries. Participants were randomly assigned to one of the three earnings releases.

While we kept the number of bullet points constant across summaries, text length could vary. Participants therefore rated summary length first in order to reduce any subconscious effects of length on subsequent judgments. Participants were not told that any of the summaries were generated automatically but they did assess whether the summary was likely to be written by management and made judgments on "Informativeness," "Readability," "Credibility," and "Overall usefulness."

As in other financial accounting studies using MTurk participants (e.g., Bonner, Clor-Proell and Koonce 2014), participants were required to pass certain screening questions to pass certain screening questions to ensure that they possess sufficient investment experience to complete the experimental task. Specifically, they were required to be over 18 years of age, to be native

English speakers, to have previous investing experience, and to be at least moderately familiar with financial disclosures (indicated by reported familiarity of 60 or higher out of 100). In total, 153 met the qualification requirements and completed the study. On completion, participants were paid $1.50 via MTurk. A mean (median) completion time of 12 (8) minutes resulted in a mean (median) hourly rate of $7.50 ($11.25).

Results are reported in the table on the next page. On average, automatic summaries of earnings releases are rated as more informative and more credible than management summaries (both $p < 0.01$), and do not differ from management summaries in terms of readability or overall usefulness (both $p > 0.10$). We also observe variation among the summarization algorithms, with LexRank, Luhn and TextRank generally getting the most favorable ratings for summaries of earnings releases. Of these, Luhn and TextRank produce by far the longest summaries. While results generally remain significant when controlling for summary length (both actual word count and perceived length), significance levels decrease for Luhn and TextRank summaries when controlling for length. This suggests that at least part of their outperformance is explained by greater length. For this reason, we exclude these two summaries from subsequent analysis, and focus instead on LexRank and other algorithms that produce summaries that are more similar in length to management summaries.

TABLE OA.1
Preliminary Test: Automatic vs. Management Summaries' Attributes

	(1) KL	(2) LEX	(3) LSA	(4) LUHN	(5) SB	(6) TR	**Average Auto**	**Average Management**
Length	33.46**	48.99***	45.41***	70.03***	34.35*	69.48***	**50.30***	**36.93**
Informativeness	63.25	67.42**	62.00	78.42***	58.10**	76.70***	**67.65***	**63.33**
Readability	71.03	73.25	70.97	69.60	69.10	66.14**	**70.01**	**71.10**
Credibility	66.12	71.37***	69.24**	76.69***	65.71	74.47***	**70.60***	**65.20**
Usefulness	64.55	69.95*	64.18	75.33**	59.54***	74.46**	**67.94**	**65.79**
Written by management	56.86	69.92***	64.39**	68.61***	53.39**	68.98***	**63.72****	**58.87**

Note: For this preliminary test, participants (N = 153) did not have access to the full text of the underlying earnings release. We used the following three earnings releases: Alibaba Q1 2016, Boeing Q2 2008, Target Q4 2013. Summary type including the management summary was manipulated within subjects and the order was randomized. For each summary, participants first rated the length of each summary. Next they rated each summary on the four attributes "Informativeness," "Readability," "Credibility," and "Overall usefulness." Finally, they assessed how likely they feel that the summary was written by the company's management. All variables are measured on a 101-point scales with appropriately labeled endpoints; higher values indicated higher levels of the measured variable. *,**,*** indicate different from management summary at $p < 0.10$, $p < 0.05$ and $p < 0.01$, respectively (all two-tailed).

Identifying Sentiment Vocabulary

In the accounting and finance literature, out of several sentiment word lists, two have been used extensively: Henry (2008), developed in the context of earnings releases, and Loughran and McDonald (2011), created in the context of 10-K filings. (For a discussion of these and other sentiment word lists (e.g., Harvard IV-4, Diction), see Loughran and McDonald (2016). To empirically motivate our choice of word list, we use regression analysis to correlate participants' "Bias" judgments, from the user evaluation test reported in the paper, with an earnings release's net sentiment score (positive word frequencies minus negative word frequencies), calculated using either the Henry (2008) or the Loughran and McDonald (2011) sentiment word list. In all regressions, we include summary and earnings release fixed effects. Finally, as a sensitivity check, we also use signed variables: specifically, the dependent (independent) variable can take three possible values: −1 if "Bias" (Net sentiment) < 0; 0 if "Bias" (Net sentiment) equals 0, +1 if "Bias" (Net sentiment) > 0. As reported in the table on the next page, the coefficients on Net sentiment are statistically significantly positive only when using the Henry (2008) word list. Using a chi-square test, we find that the difference in coefficients is statistically significant at 6% or better (two-tailed).

Given these results, in the remainder of the paper, we utilize Henry's (2008) word list to measure positive and negative tone.

TABLE OA.2
Identifying Sentiment Vocabulary

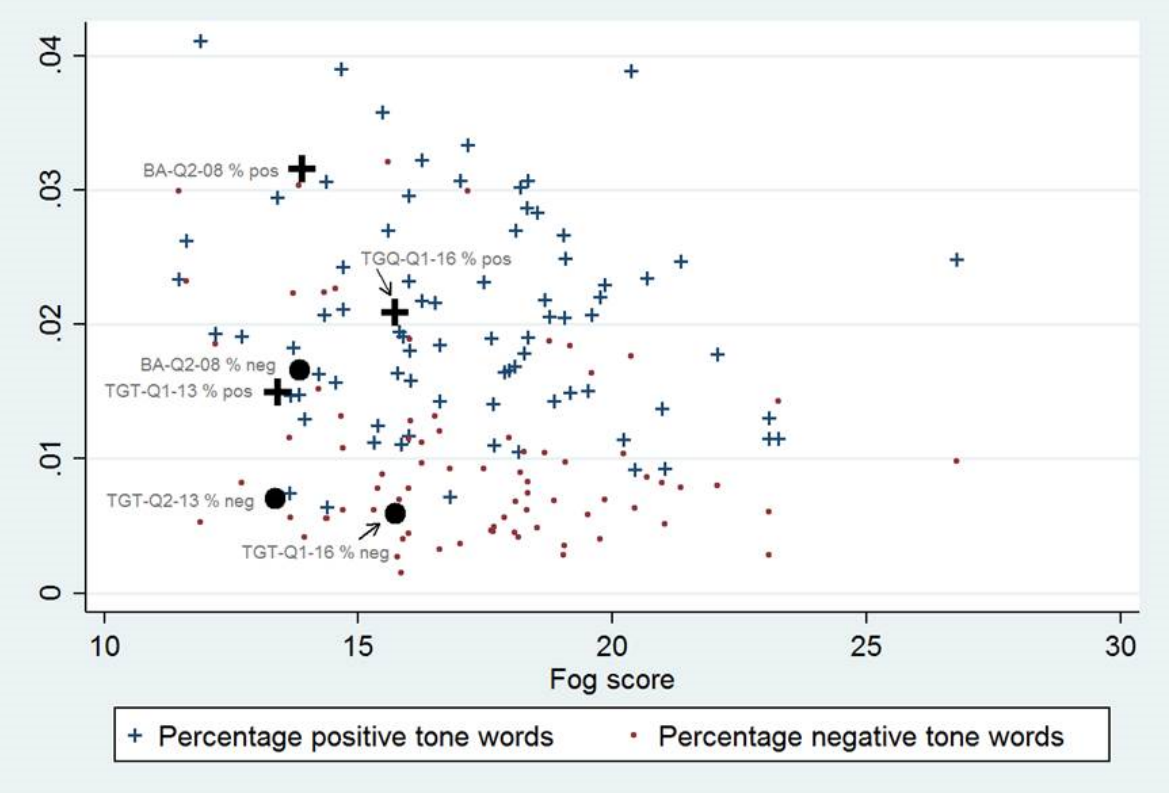
	"Bias"		Signed "Bias" { Neg. [−1], Neut. [0], Pos. [+1] }	
	(1) Henry (2008)	(2) Loughran and McDonald (2011)	(3) Henry (2008)	(4) Loughran and McDonald (2011)
Net sentiment	1.9878***	0.5057		
	(3.37)	(0.94)		
Signed net sentiment { Neg. [−1], Neut. [0], Pos. [+1] }			0.3284***	0.0394
			(3.18)	(0.68)
Test coefficient equality	(1) = (2) $\chi^2 = 3.50$*		(3) = (4) $\chi^2 = 6.02$**	
Summary fixed effects	Included	Included	Included	Included
Earnings release fixed effects	Included	Included	Included	Included
Number of observations	980	980	980	980
R-squared (%)	3.85	1.92	5.11	3.23

This table presents estimates from ordinary least-squares linear regressions of "Bias" (columns 1-2) and Signed "Bias" (columns 3-4) on Net sentiment and Signed net sentiment, respectively, with robust standard errors. "Bias" judgments were made on a 101-point scale with endpoints of –50 ("Summary makes [Company] look worse") and +50 ("Summary makes [Company] look better"). Signed "Bias" can take three possible values: −1 if "Bias" < 0; 0 if "Bias" equals 0, +1 if "Bias" > 0. Net sentiment is calculated as number of positive tone words -/- number of negative tone words. Signed net sentiment can take three possible values: −1 if Net sentiment < 0; 0 if Net sentiment equals 0, +1 if Net sentiment > 0. Two frequently used sentiment words lists are used: Loughran and McDonald (2011) and Henry (2008). We include summary and earnings release fixed effects in all regressions. The table also reports the chi-square test for equality of coefficients. We use ***, **, and * to denote that the coefficient estimate is statistically different from zero at the 1%, 5%, and 10% levels (two-tailed), respectively.

Sample Descriptives

Figure OA.1 plots % positive (plus-sign) and % negative (circle) tone words (i.e. number of positive and negative tone words, respectively, scaled by total number of words) on the Y-axis and readability (Gunning-Fog score) on the X-axis for N = 78 earnings releases. (See Table 3 in the paper.) The three cases from the paper (Boeing, Target, and HSQ) are also positioned in the graph. The average Fog score is 17.1 (untabulated), which corresponds closely with a (large sample) average of 16.8 documented in Bozanic, Roulstone, and Van Buskirk (2014), which suggests that the average earnings release requires a college education to be understood.

FIGURE OA.1
Distribution of Sample Earnings Releases: Textual Sentiment and Readability



'TGT-Q1-13', 'BA-Q2-08', and 'TGT-Q1-16' represent Target's Q1-2013 earnings release, Boeing's Q2-2008, and Target's Q1-2016 earnings release (= HSQ's earnings release used in the experiment), respectively.

This graphical analysis shows that, as regards readability, all three cases have a college-reading level (i.e. Fog score >= 13). As for net textual sentiment (calculated as # positive tone words - # negative tone words, scaled by total words), all cases fall reasonably close (i.e. within one-standard-deviation range) to the sample mean of 1%. Thus, increasing the potential generalizability of our results.

REFERENCES

Bonner, S.E., Clor-Proell, S.M., & Koonce, L. (2014). Mental accounting and disaggregation based on the sign and relative magnitude of income statement items. *The Accounting Review*, 89, 2087-2114.

Bozanic, Z., Roulstone, D., & Van Buskirk, A. (2014). *Attributes of informative disclosures.* Working paper.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Seventh International World-Wide Web Conference*, Brisbane, Australia.

Erkan, G., & Radev, D. (2004). LexRank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22, 1-23.

Gong Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. *Proceedings of SIGIR*, 19-25.

Haghighi, A., & Vanderwende, L. (2009). Exploring content models for multi-document summarization. *Proceedings of Human Language Technologies*, 362-370.

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *Journal of Business Communications*, 45, 363–407.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66, 35-65.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54, 1187-1230.

Luhn, H.P. (1958). The automatic creation of literature abstracts, *IBM Journal*, April, 159-165.

Mani, I. (2001). *Automatic summarization*, Natural Language Processing. John Benjamins Publishing Co.

Mani I., & Maybury, M. (1999). *Advances in automatic text summarization.* MIT Press.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. UNT Digital Library. Retrieved from http://digital.library.unt.edu/ark:/67531/metadc30962/. Accessed December 19, 2016.

Nenkova, A. & McKeown, K. (2011). Automatic summarization. *Foundations and Trends in Information Retrieval*, 5, 103-233.

Radev, D., Hovy, E., & McKeown, K. (2002). Introduction to special issue on summarization. *Computational Linguistics*, 28, 399-408.

Steinberger, J., & Jezek, K. (2004). Using latent semantic analysis in text summarization and summary evaluation. *Proceedings of ISIM*, 93-100.

Torres-Moreno, J-M. (2014). *Automatic text summarization*. Hoboken, NJ: Wiley.

Van der Kaa, H., & Krahmer, E. (2014). Journalist versus news consumer: The perceived credibility of machine written news. In: *Proceedings of the Computation + Journalism Conference*. New York.

Vanderwende, L., Suzuki, H., Brockett, C., & Nenkova, A. (2007). Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing and Management*, Special Issue on Summarization, 43, 1606-1618.