RFS Advance Access published April 29, 2010

The Information Content of IPO Prospectuses

Kathleen Weiss Hanley
Federal Reserve Board of Governors, Division of Research and Statistics

Gerard Hoberg
University of Maryland, Robert H. Smith School of Business, College Park

Using word content analysis, we decompose information in the initial public offering prospectus into its standard and informative components. Greater informative content, as a proxy for premarket due diligence, results in more accurate offer prices and less underpricing, because it decreases the issuing firm's reliance on bookbuilding to price the issue. The opposite is true for standard content. Greater content from high reputation underwriters and issuing firm managers, through Management's Discussion and Analysis, contribute to the informativeness of the prospectus. Our results suggest that premarket due diligence and disclosure by underwriters and issuers can serve as a substitute for costly bookbuilding. (*JEL* G14, G24, G30, G32)

Althoughthere exists a substantial body of literature on the initial public offering (IPO) process and the determinants of IPO pricing, unresolved questions remain on how IPOs are priced. Traditional views of IPO pricing have focused primarily on the role of bookbuilding in providing valuable information for IPO pricing (Benveniste and Spindt 1989; Spatt and Srivastava 1991; Cornelli and Goldreich 2003). In this article, we examine whether there exists a trade-off between increasing the accuracy of IPO pricing in the premarket using fundamental analysis (Kim and Ritter 1999) and costly information gathered from investors during bookbuilding.

This trade-off is as follows. The underwriter and issuing firm could expend resources in the premarket (before filing the initial prospectus with the Securities and Exchange Commission [SEC]) to collect information about the offer,

We are especially grateful to Michael Weisbach (the editor) and two anonymous referees for very helpful suggestions. We thank Scott Bauguess, Wolfgang Bessler, Valentin Dimitrov, Laura Field, Robert Hauswald, Ayla Kayhan, Josh Lerner, Lubo Litmov, Tim Loughran, Michelle Lowry, Tom Noe, Jennifer Marietta-Westberg, Stas Nikolova, Ann Sherman, Chester Spatt, Paul Tetlock, and Josh White as well as seminar participants at American University, Pennsylvania State University, the Securities and Exchange Commission, the University of Missouri, the University of Notre Dame, the University of Western Ontario, Vanderbilt University, Washington University, the EFM IPO Symposium, the 19th Annual Conference on Financial Economics and Accounting, and the 2010 AFA Annual meeting. The ideas and opinions expressed in this article are the authors' and should not be interpreted as reflecting the views of either the Board of Governors of the Federal Reserve System or its staff. All errors are the authors' alone. Send correspondence to Kathleen Weiss Hanley, Federal Reserve Board of Governors, Division of Research and Statistics, Risk Analysis Section (Mail Stop 91), 20th and C Streets NW, Washington, DC 20551; telephone: (202) 973–7324. E-mail: kathleen.hanley@frb.gov.

doi:10.1093/rfs/hhq024

which is then used to set the initial offer price or range. The accuracy and substance of this initial price will be believed by investors, however, only if it is accompanied by informative disclosure in the prospectus and during the road show. Investing in information production in the premarket, however, could be costly, because it involves substantial effort on the part of the underwriter, the issuing firm, and both their legal counsels to collect information through due diligence. Furthermore, the disclosure of this information may reveal valuable strategic or proprietary information to rivals (Bhattacharya and Ritter 1983; Darrough and Stoughton 1990; Bhattacharya and Chiesa 1995).

Thus, it could be the case that the issuer and the underwriter conclude that premarket information production for the purposes of pricing the issue is too expensive and may instead choose to rely on information produced by investors during the bookbuilding period. In the extreme case where premarket information gathering is prohibitively expensive, the underwriter and issuer could simply disclose the minimum necessary for regulatory and liability reasons and essentially let investors value the IPO during bookbuilding. However, the choice to use information generated from investors can also be costly for issuers, because investors must be compensated for providing an accurate assessment of value through higher initial returns (Hanley 1993; Sherman and Titman 2002). The decision on which method to use to gather information, enhanced premarket due diligence or bookbuilding, hinges on the relative costs and benefits of each.

The focus of this article is to examine whether such a trade-off exists and to measure its impact on pricing. Although the amount of premarket information production is not directly observable, we use a unique methodology that decomposes the initial IPO prospectus into its standard content and informative content components. Standard disclosure is defined as the exposure to information in an IPO prospectus that is already contained in both recent and past industry IPOs, while informative content is the disclosure in the prospectus (residual) not explained by these two sources. We use the relative exposure to each type of content to measure the degree of premarket effort expended by underwriters and issuers.[1]

Greater information produced during the premarket will produce content that is unique to a given IPO and informative to readers of the prospectus. More informative content should result in more accurate initial offer prices relative to final offer prices before bookbuilding begins and, thus, reduce the need for information generated during bookbuilding. We predict that offers with greater informative content will have smaller absolute changes in offer prices (relative to the initial filing price estimate) and lower underpricing. If, instead, issuers

[1] Note that informative content is the absolute value of the residuals of a regression estimating standard content. This means that informative content is not simply the mirror image of standard content but represents deviations both positive and negative from the estimated exposure to standard content. This issue is discussed in further detail in Section 4.1.

and underwriters choose to invest less in premarket due diligence, then disclosure will have a higher exposure to standard rather than informative content, as more of the prospectus is likely to be "copied" from other sources such as recent and past industry IPOs. Since these issuers will have more price discovery determined during bookbuilding, greater standard content will result in greater offer price changes and higher underpricing as compensation to investors for revealing information.[2]

Our findings support these predictions. The greater the informative content in the prospectus, the better the pricing accuracy. This improved accuracy is economically meaningful, as a 1 standard deviation (SD) increase in informative content is associated with an approximately 3% reduction in the absolute value of the change in offer price and a 2% change in the level of the change in offer price. We also find a substantial reduction of 8% in the level of underpricing, which is consistent with the substitution effect of premarket due diligence for costly bookbuilding.

The opposite effect holds true for standard content. Although the absolute change in the offer price is economically insignificant, we do find that a 1 SD increase in standard content equates to an approximately 1.5% increase in the level of the offer price, with a corresponding 4% increase in initial returns. These results suggest that if information production is costly to investors (Sherman and Titman 2002), issuers seeking to have their IPO priced using bookbuilding might have an incentive to lowball the initial offer price. By doing so, issuers and underwriters signal their willingness to pay information rents to investors and induce them to produce more information during bookbuilding.

By parsing the prospectus into its four most important sections—the Prospectus Summary, Risk Factors, Use of Proceeds, and Management's Discussion and Analysis (MD&A)—we find the strongest association between informative and standard content and pricing to be in MD&A. Although prior work has generally focused on the role of the underwriter in information production, our results indicate that issuing firm managers, through MD&A, might perform a surprisingly integral role in the information generation process. This is especially true for IPOs that are more likely to be priced using premarket due diligence rather than bookbuilding and is indicative that managerial involvement in this process adds value. A 1 SD increase in informative content in this section is related to a 2.6% decline in the absolute change in offer price, a 4.5% decline in the level of the change in offer price, and a 12% decline in initial returns. These findings support our contention that it is both the issuing firm and the underwriter who jointly contribute to price

[2] Alternatively, the amount of standard content could be viewed as a measure of how similar an IPO is relative to its counterparts. Here, an IPO with a high degree of standard content could be very similar to other recent IPOs and should be easier for both the underwriter and investors to price. Such an IPO would have less information generated during bookbuilding due to its low valuation uncertainty, and we should observe a more accurate initial price and lower initial returns. Our results do not support this alternative, as the presence of more standard content uniformly results in less accurate initial prices and also greater underpricing, which is consistent with heterogeneous effort expended on due diligence in the premarket.

discovery. It also confirms the Jenkinson and Jones (2009) finding that many fund managers find one-on-one meetings with management to be useful in terms of forming a view on valuation.

Since the literature on disclosure predicts similar relationships as those above (see Verrecchia 2001; Dye 2001; Healy and Palepu 2001), it is natural to ask if our strong pricing results are due simply to differences in the amount of disclosure of known information or are indeed linked to heterogeneous levels of effort being expended on due diligence.[3] In order to determine whether differences exist in the amount of effort expended by issuers, we examine the relation between content type and issuer expenses. Greater premarket due diligence should generate larger fees to compensate lawyers, accountants, and investment bankers for their additional time; however, these fees are not likely to be significantly related to whether or not an issuer discloses certain known information.

Indeed, we find that greater informative content is associated with higher issuer expenses, particularly those related to legal and auditing fees. The component of the gross spread most likely to reflect greater due diligence, on the part of the underwriter the management fee, is also significantly related to the amount of informative content. Standard content, on the other hand, is related to only the selling fee component of the gross spread, which is consistent with less premarket due diligence, requiring greater selling effort (substitution toward more information from bookbuilding) during the offering process. These findings are consistent with heterogeneous differences in premarket due diligence effort.

Since underwriter content is an important source of prospectus content in general and the focus of much of the IPO literature in particular, we examine whether unique underwriter content contributes to pricing accuracy. We show that greater unique underwriter content is associated with greater pricing accuracy over and above standard and informative content. However, this increase in accuracy is associated with unique content from only higher-reputation underwriters, as unique content from lower-reputation underwriters has no effect on pricing. One possible interpretation is that greater underwriter content is a proxy for underwriter-specific effort in drafting the prospectus (which can translate to more information relevant to pricing), consistent with the literature on underwriter certification.

Finally, we explore the type of informative content that is most significantly related to pricing and find that content directly related to inputs into valuation models most likely used by practitioners seems to matter most. For example, we find that greater disclosure related to word lists from areas including accounting, corporate strategy, valuation, product markets, and corporate governance, is associated in all of these aspects with significant reductions in the

[3] Note that the effort story posited above and the alternative addressed here are not mutually exclusive. Disclosure strategy may also affect the amount of due diligence and vice versa.

change in offer price and underpricing. In contrast, legal text matters little, and marketing text runs in the opposite direction by decreasing pricing accuracy.

The tone of the prospectus has a strong relation to pricing for only the Risk Factors section, where more positive text is associated with increased pricing accuracy. This result is consistent with this section's role of mitigating liability risk. Since the underwriter and issuer are liable, both legally and reputationally, for any misstatements in the prospectus, a net positive tone sends a strong signal to investors regarding the riskiness and valuation of the issue, which is associated with an increase in pricing accuracy.

Our work builds on prior literature that examines disclosure in the context of IPOs. For example, Beatty and Ritter (1986) present evidence that greater information in the Use of Proceeds increases underpricing. In contrast, Leone, Rock, and Willenborg (2007) and Ljungqvist and Wilhelm (2003) find that firms that are more (less) specific in their disclosure of the uses have lower (higher) underpricing. Other papers have examined the relation between the size of the Risk Factors section and pricing. Beatty and Welch (1996) and Arnold, Fishe, and North (2008) examine the Risk Factors section of the prospectus and find that greater disclosure in this section is associated with higher initial returns. Guo, Lev, and Zhou (2004) focus on product-related disclosures in the prospectus by firms in the biotechnology industry. They find a negative relation between the extent of disclosure and the bid–ask spread but do not examine if there is a link to IPO underpricing.

Recent papers on media and company press releases have also highlighted the importance of disclosure for IPO pricing. Schrand and Verrecchia (2005) find that greater pre-IPO disclosure frequency reduces underpricing, while Cook, Kieschnick, and Van Ness (2006) present evidence that the greater the number of news articles prior to going public, the larger the price revision and underpricing. Liu, Sherman, and Zhang (2007) argue that the effect of pre-IPO media coverage differs when positive and negative information is revealed during bookbuilding.

Finally, our article reflects a growing interest in the use of word content analysis to analyze the informativeness of written disclosure and media coverage. In the context of managing litigation risk, Nelson and Pritchard (2008) and Mohan (2007) find that certain word usage is related to the probability of being sued. Hoberg and Phillips (2008) use text similarity analysis to test theories of merger incidence and outcomes. Loughran and McDonald (2008) show that firms using plain English have greater small investor participation and shareholder-friendly corporate governance. In other contexts, papers such as those of Tetlock (2007), Tetlock, Saar-Tsechanksy, and Macskassy (2008), Li (2006b), and Boukus and Rosenberg (2006) find word content to be informative in predicting stock price movements.

The remainder of the article is organized as follows: Section 1 includes a discussion of the mechanics and theory of IPO pricing. The data, methodology, and summary statistics are presented in Section 2. The sources of prospectus

content are explored in Section 3. The method to decompose the prospectus text into standard and informative content, as well as tests relating these content measures to pricing and expenses, is in Section 4. Section 5 considers unique underwriter content, and Section 6 examines the specific types of content contained in each section as well as the tone of the corresponding text. The article concludes in Section 7.

1. The IPO Pricing Process

After the issuer chooses an underwriter(s), there are three steps in the pricing of an IPO. First, the underwriter and the issuing firm conduct due diligence, draft an initial prospectus that is filed with the SEC, and set the initial offer price. We define the initial offer price as the midpoint of the initial offer price range. Second, a final offer price is specified using information gathered from investors during bookbuilding. If no new information is revealed, then the final offer price should be equal to the initial offer price.[4] Finally, a market price is established once trading begins and the initial return or underpricing is determined.[5]

Much of the literature on IPOs has focused on how final offer prices are set with, an emphasis on the role of bookbuilding in pricing an issue. However, information gathered from investors during bookbuilding is an expensive mechanism to price an offer (Hanley 1993). There have been few efforts to determine whether there exists an alternative mechanism to bookbuilding (other than auctions) that could reduce this high cost.

With the exception of Kim and Ritter (1999) and Lowry and Schwert (2002), there has been little research on how issuers and underwriters determine the initial value of the IPO, and most papers assume that the preliminary offer price is an unbiased predictor of the expected offer price given fundamental information about the firm. Substantial resources, however, are expended on due diligence by the underwriter, the issuing firm, and their legal counsel to gather information about the firm. While some of this expenditure is due to regulatory or liability concerns, it is plausible that greater effort expended in the premarket to acquire information about both the issuing company and its competitors may lead to more accurate initial offer prices relative to final offer prices. In this article, we propose that issuers and underwriters can choose to engage in price discovery in the premarket (prior to the filing of the initial prospectus

[4] There may be reasons why this relationship may not hold in practice. For example, underwriters and issuers may set the initial offer price lower than expected in order to provide an incentive for investors to invest in information production. This alternative is discussed later in the article.

[5] For the purposes of this discussion, we assume that the market price is invariant to how the issue is priced at the time of the offer. This assumption may not hold, as greater information production and disclosure may have an impact on the aftermarket valuation of the issue by investors.

with the SEC) or during bookbuilding.[6] If the issuer and underwriter choose to have more accurate pricing in the premarket, they will expend greater effort in acquiring information through enhanced due diligence about the issuing firm and its competitors. The accuracy of this initial price will be believed by investors, however, only if it is accompanied by credible disclosure in the initial prospectus and during the road show.[7]

The benefit of increased information acquisition during the premarket is that the initial offer price will be a more accurate assessment of both the final offer and aftermarket trading prices. Because less information will be gathered from investors during bookbuilding, issuers conducting premarket due diligence benefit from both lower price changes and lower compensation to informed investors in the form of underpricing (Benveniste and Spindt 1989). This benefit, however, is offset by the potential cost of revealing proprietary information to rivals. Further, enhanced due diligence may become too expensive due to higher legal fees and underwriter compensation. This is particularly true if the underwriter's compensation cannot be raised high enough to compensate for the additional effort (see Chen and Ritter 2000 for evidence on limits on underwriter compensation).

If the cost of gathering information in the premarket becomes too expensive relative to the benefit of more accurate pricing and lower initial returns, the issuing firm and the underwriter may choose, instead, to price the issue using information gathered from investors during bookbuilding. In this case, the issuing firm and the underwriter do not expend resources to set accurate initial offer prices that fully incorporate available information but, instead, use information produced during bookbuilding to price the issue.[8] While this reduces the cost of premarket information production, investors must be rewarded through both increased allocation and initial returns for truthfully revealing their valuation of the issue. Many estimates of the cost of rewarding investors for providing this information suggest that it is high.[9] During this study's sample period, issues that priced above their initial offer price had average underpricing of 66%.

The final choice between pricing the issue using premarket information production or information generated during bookbuilding may also reflect a tension between underwriters and issuing firms. Underwriters may have an

[6] In reality, price discovery in the premarket or during bookbuilding is not an either-or decision but is more likely a continuum.

[7] There are a number of consequences to including false or misleading information in the prospectus. First, when the true value of the information is eventually revealed in the aftermarket, share prices will decline and IPO participants might be sued. Second, obfuscation or false statements in the prospectus can damage reputational capital, particularly underwriters and lawyers, over and above the value of legal damages. Finally, the SEC, which reviews IPO filings, may also scrutinize and comment on the inclusion of useless or misleading statements (see Ertimur and Nondorf 2009).

[8] This may explain the puzzling finding of Lowry and Schwert (2004) "that public information released before the filing is not always fully incorporated into the filing range."

[9] Evidence on this proposed relation is contradictory (see Jenkinson and Jones 2004).

incentive to limit the amount of effort expended in the premarket especially if underwriting fees are capped.[10] Since they do not bear the consequences of the cost of bookbuilding and may also benefit from greater underpricing, underwriters may view information gathering during the bookbuilding process as more cost effective and advantageous. If the cost of revealing proprietary information is not prohibitive, some issuers will prefer to have more accurate pricing and lower initial returns through investment in premarket information production. Loughran and Ritter's (2002) theory of relative bargaining power would suggest that more powerful issuers will be more successful in convincing underwriters to expend effort in the premarket even if underwriting fees are not sufficient to compensate them.

Resource constraints during hot issue markets may also play a role. Khanna, Noe, and Sonti (2008) model the effect of a tight labor market in the investment banking industry on aggregate underpricing. In their model, increased demand for the underwriter's services coupled with constraints in the labor market reduces the ability to produce information while at the same time increasing the cost, which then increases underpricing. Their model suggests (but does not explicitly address) that the underwriter may substitute in-house or premarket information production with information gathered during bookbuilding when it is more efficient to do so.

The outcome of greater information production in the premarket should be more informative content in the preliminary prospectus. Alternatively, when issuers and underwriters choose to use bookbuilding, the initial prospectus is more likely to be composed of information that is publicly available from recent and past industry IPOs and will have a higher degree of standard content.[11] By decomposing the information in the initial prospectus into standard and informative content, we are able to proxy for the relative amount of effort expended in the premarket to price the issue and to assess its impact on pricing. We suggest that disclosure in the initial prospectus is representative of the amount of effort expended in premarket due diligence.[12]

The following empirical predictions emerge from this view of IPO pricing:

1. Issuers and underwriters who choose to invest in greater information production in the premarket are predicted to have more informative content in the initial prospectus and less need to rely on bookbuilding in pricing the issue. We, therefore, predict that greater informative content should be associated with more accurate initial offer prices relative to final offer prices (smaller absolute offer price changes). Underpricing will also be lower

[10] This discussion assumes that the underwriter provides enough premarket disclosure to limit its liability risk.

[11] Some standard content may be included to satisfy regulatory requirements on sufficient information in the prospectus. In addition, the trade-off between premarket information gathering and bookbuilding, as well as the choice of legal counsel and/or underwriter, may be influenced by the perceived response by regulatory authorities to prospectus content.

[12] We examine whether this is indeed the case in Section 4.3.

the greater the informative content is, since issuers and underwriters need not compensate investors for providing information during bookbuilding (Benveniste and Spindt 1989).

2. Issuers and underwriters who choose to expend fewer resources on premarket due diligence will have more information in the initial prospectus that is standard or copied from other sources and will rely heavily on bookbuilding for price discovery. Conversely to the case of higher informative content, we expect that greater standard content should result in both larger absolute price changes (more information generated during bookbuilding) and higher initial returns as compensation to investors for revealing information.

The above discussion is based on the assumption, used by many studies in the existing IPO literature, that the initial IPO price is a fair value estimate of the final offer price given all information known at the time of initial filing. We acknowledge, however, that this may not be the case. Because there are differences in the amount of effort expended on premarket due diligence, the need for information generated during bookbuilding is not homogeneous across all issuers. Thus, if information is costly for investors to produce (Sherman and Titman 2002), it is not always efficient for investors to produce the same amount of information for every offer. This begs the question of how issuers can credibly convey the level of information production for which they are willing to pay. One possible mechanism is to set the initial offer price below the fair value given the information known at the time of filing.

Issuers seeking more information production can lowball the initial price, therefore signaling the need for information production and promising larger economic rents through the well-known partial adjustment mechanism. As stated by Sherman and Titman (2002), "If the pricing policy provides a sufficient incentive for investors to collect information, they will also have sufficient incentive to reveal the information." Under this scenario, standard content should be positively correlated not only with absolute price changes but also with higher levels of offer price changes. Future theoretical studies examining how the initial price estimate is formed are needed to further inform this debate.

2. Data and Methodology

2.1 Sample and word vector construction

We obtain our initial list and characteristics of all U.S. IPOs issued between January 1, 1996 and October 31, 2005, from the Securities Data Company U.S. New Issues Database. We eliminate American Depository Receipts, unit issues, Real Estate Investment Trusts (REITs), closed-end funds, financial firms, and firms with offer prices less than $5. A Center for Research in Security Prices (CRSP) PERMNO must also be available for an observation to remain

in the sample, and the IPO must also have a valid founding date, as identified in the Field–Ritter dataset, as used in Field and Karpoff (2002) and Loughran and Ritter (2004).[13] These initial exclusions reduce the sample to 2,112 IPOs.

For each IPO passing these initial screens, we use a Web crawling algorithm to download the initial prospectus. In order for an IPO to remain in our sample, it must have an SEC Edgar filing available online, and the online document must also be machine readable. In order to satisfy our definition of machine readable, a Table of Contents pagination algorithm must be able to detect and accurately identify the start and end of the four key sections of the prospectus. These sections are the "Prospectus Summary," "Risk Factors," "Use of Proceeds," and "Management's Discussion and Analysis." This additional screen eliminates sixty-nine IPOs, leaving us with 2,043 machine-readable IPOs.[14] Because these sixty-nine IPOs are a small fraction of our sample and because most are also small firms that file using an SB-2 (larger firms generally file an S-1), we believe that our results are likely to be unaffected by their exclusion.

Our estimation of standard content incorporates information from prior IPOs. In order to have sufficient data for the estimation, the sample is further restricted to IPOs that were issued on or after August 1, 1996. IPOs issued prior to that date are used to compute starting values for information on recent and past industry IPOs. In order to estimate our model of informative and standard content as discussed in Section 4.1, we include only IPOs that have prospectus data from (i) at least one other IPO that was filed ninety days prior to the current IPO's filing date and (ii) at least one other IPO in the same Fama–French forty-eight industry code as the current IPO that was filed at least ninety-one days prior to but no later than one year before the current IPO's filing date. This requirement reduces our sample to one thousand seven hundred fifty IPOs.

Our algorithm to read each prospectus is written in a combination of PERL and APL. Once a document is downloaded and paginated, our algorithm's next step is to purge the document of attachments, headers, and exhibits so that we can focus on the prospectus itself. This is achieved using a three-pronged approach that ensures a very high degree of accuracy: (i) We use the pagination implied by the Table of Contents to identify the beginning and end of the document, (ii) we examine the placement of the "additional information" statement and the placement of accounting statements (exhibits) to confirm accuracy,[15]

[13] We thank Jay Ritter for generously providing the database of IPO founding dates on his Web site.

[14] A significant amount of work has been done to maximize the fraction of prospectuses that are deemed machine readable. This includes hand checking each prospectus failing our machine-readability condition to determine if our document pagination algorithm can be improved via exception handling. An example of an exception is that some filings have slight variations in the section names which we list. For example, the Prospectus Summary is occasionally called "Summary." The sixty-nine IPOs failing machine readability generally lack pagination or may even lack a Table of Contents.

[15] The overwhelming majority of prospectuses filed in our sample have a statement indicating where investors can find additional information toward the end of the prospectus document.

and (iii) we hand check the algorithm's accuracy for most documents and include exception handling where necessary.

For each IPO i, we store the text of the prospectus in a word vector, which we define as $words_{tot,i}$. We also store the text from each of the four sections in separate word vectors, which we define as $words_{ps,i}$ (Prospectus Summary), $words_{rf,i}$ (Risk Factors), $words_{use,i}$ (Use of Proceeds), and $words_{mda,i}$ (Management Discussion and Analysis). Our words are based on word roots rather than actual words and exclude certain types of words such as common words and/or articles (for additional information on the word vector construction, see Appendix A). For every IPO, we have one such vector for the document as a whole and each of the four sections. Note that all word vectors for both the entire document and each of the four sections have the same length (5,803), as they are based on the same global word list of 5,803 word roots. Each element of the vector is populated by the count of the number of times the word is used in the given document.

As an example, consider a simple universe of two prospectuses, one with the content "they sell potatoes and they sell corn" and one with the content "they sold knives." Discarding articles, conjunctions, and pronouns (*the*, *and*, and *they*), there are four word roots in the union of both documents: *sell*, *potato*, *corn*, and *knife*. In our example, we have:

$$words_{tot,1} = \{2, 1, 1, 0\} \qquad \text{and} \qquad words_{tot,2} = \{1, 0, 0, 1\}.$$

Note that when the underlying document is larger, these word vectors are populated with more words. Hence, these vectors measure the "total amount" of information in the document.

We then normalize the raw word vector $words_{tot,i}$ by the total number of root words used in the document and define this as $norm_{tot,i}$.[16] Normalized word vectors have elements that sum to 1 and do not sum differently when a document is larger.[17] Therefore, the normalization of the vectors noted above would be:

$$norm_{tot,1} = \{0.50, 0.25, 0.25, 0\} \qquad \text{and} \qquad norm_{tot,2} = \{0.50, 0, 0, 0.50\}.$$

Later in Section 4, we decompose the total amount of disclosure into standard and informative content.

[16] Our results are robust to whether the vector is scaled by total document size or by root word size and to using the cosine method to normalize word vectors, where vectors are normalized to have a length of 1 rather than a sum of 1. We focus on the current normalization, as it permits a simpler decomposition into standard and informative content.

[17] It is an open question as to whether disclosure should be measured as raw or scaled word count. Each has its benefits and attendant costs. Raw word counts measure the magnitude of disclosure but may also capture a size effect. Scaled word counts eliminate a size bias but cannot capture raw quantity. We present results based on scaled word counts because the direct link between raw word counts and document size induces a high degree of correlation between content types (nearly 80%), as larger documents have more of both types. In contrast, our scaled content types are less than 4% negatively correlated. Although we do not report them, our results are strongest when we use raw content variables. However, given their high correlation, we present the scaled results in order to be conservative.

2.2 Other control variables

We also compute a number of pricing variables that are common to the existing IPO literature:

$$\Delta P = \frac{P_{ipo} - P_{mid}}{P_{mid}} \qquad \text{and} \qquad IR = \frac{P_{mkt} - P_{ipo}}{P_{ipo}}.$$

P_{mid}, P_{ipo}, and P_{mkt} are the filing date midpoint, the IPO price, and the aftermarket trading price, respectively. ΔP is the offer price adjustment from the filing date to the IPO date, and IR (initial return) is the market's price adjustment from P_{ipo} to P_{mkt}. Investors who purchase shares at the IPO price, P_{ipo}, can realize returns equal to IR by selling their shares at the closing price on the first day of public trading.

We also account for the following variables identified in the existing IPO literature to control for characteristics specific to the IPO:

Firm Age: IPO year minus the firm's founding date, where founding dates are obtained from the Field–Ritter dataset, as used in Field and Karpoff (2002) and Loughran and Ritter (2004).

Lead UW $ Market Share: Lead underwriter's dollar market share in the past calendar year as calculated by Megginson and Weiss (1991).

Law $ Market Share: This variable is calculated as the dollar market share in the past calendar year, and a separate variable is constructed for the lead underwriter's legal counsel and the issuer firm's legal counsel.

VC Dummy: Dummy variable equal to 1 if the firm is venture capital (VC) backed and 0 otherwise, as in Barry, Muscarella, Peavy, and Vetsuypens (1990).

Nasdaq Return: The NASDAQ return measured over the thirty trading days preceding the filing date. Logue (1973) first examined whether past market returns can predict future underpricing, and a variant of this measure has been used more recently by Loughran and Ritter (2002) and Lowry and Schwert (2004).

IPO Size: The natural logarithm of the original filing amount.

Tech Dummy: Dummy variable equal to 1 if a firm resides in a technology industry as identified in Loughran and Ritter (2004).

2.3 Summary statistics

Table 1 presents summary statistics for the sample of one thousand seven hundred fifty IPOs. Panel A has information on the price variables, and our sample is similar to other studies that include the bubble period of 1999 and 2000. On average, this sample of IPOs has an average initial return of 36.4% with a much lower median of 13.4%. The average change in the offer price from the first initial price range midpoint to the final offer price is 4.3%, and the average absolute price change is 19.3%.

Table 1
Summary statistics

	Mean	Std. dev.	Minimum	Median	Maximum
Panel A: Price variables					
Price adjustment (ΔP)	0.043	0.283	−0.984	0.000	2.200
Absolute price adjustment ($\lvert\Delta P\rvert$)	0.193	0.211	0.000	0.133	2.200
Initial return (IR)	0.364	0.691	−0.399	0.134	6.267
Panel B: IPO characteristics					
IPO size at filing	194.6	1245.1	2.8	57.9	46926.1
Gross proceeds	116.2	349.3	2.3	56.0	8680.0
Firm age	13.022	18.989	0.000	7.000	165.000
Tech dummy	0.470	0.499	0.000	0.000	1.000
VC dummy	0.490	0.500	0.000	0.000	1.000
Lead UW $ market share	0.028	0.025	0.000	0.022	0.147
UW law $ market share	0.023	0.033	0.000	0.010	0.216
Issuer law $ market share	0.012	0.022	0.000	0.004	0.177
Auditor $ market share	0.161	0.081	0.000	0.167	0.557
Pre-file Nasdaq return	0.049	0.092	−0.260	0.054	0.350
Panel C: Prospectus text					
Total words in document	33983.7	10831.8	14655.0	31965.5	119300
Prospectus Summary/total document	0.058	0.024	0.016	0.052	0.323
Risk Factors/total document	0.183	0.050	0.050	0.184	0.442
Use of Proceeds/total document	0.009	0.005	0.000	0.008	0.042
MD&A/total document	0.130	0.044	0.022	0.129	0.480
All four sections	0.381	0.057	0.203	0.380	0.789
Panel D: Word vectors (word roots only)					
Total document ($words_{tot,i}$)	9713.29	3200.61	4338.00	9147	35942
Prospectus Summary ($words_{ps,i}$)	608.66	362.36	126.00	505	3961
Risk Factors ($words_{rf,i}$)	1767.13	674.83	510.00	1704	5312
Use of Proceeds ($words_{use,i}$)	72.54	37.02	2.00	65	277
MD&A ($words_{mda,i}$)	1172.93	661.55	143.00	1044	8724

Summary statistics are reported for one thousand seven hundred fifty IPOs issued in the United States from August 1996 to October 2005 excluding firms with an issue price less than $5, ADRs, financial firms, unit IPOs, dual class IPOs, and REITs. Price adjustment (ΔP) is the return from the filing date midpoint to the IPO offer price, and Absolute price adjustment (abs(ΔP)) is the absolute value of the change in offer price. Initial return (IR) is the actual return from the IPO offer price to the first CRSP reported closing price. The IPO size at filing is the original filing amount, and the Gross proceeds are the amount actually offered. Firm age is the IPO year minus the firm's founding date, where founding dates are obtained from the Field–Ritter dataset, as used in Field and Karpoff (2002) and Loughran and Ritter (2004). The Tech dummy is equal to 1 if a firm resides in a technology industry as identified in Loughran and Ritter (2004). The VC dummy is equal to 1 if a firm is VC financed. Lead UW $ market share is the lead underwriter's dollar market share in the past calendar year. UW law $ market share is the underwriting firm's legal counsel's dollar market share in the past calendar year. Issuer law $ market share is the issuer firm's legal counsel's dollar market share in the past calendar year. Auditor $ market share is the auditor's dollar market share in the past calendar year. Pre-file Nasdaq return is the NASDAQ return for the thirty trading days preceding the filing date. The Word Vector statistics are the number of root words in each section after applying filters to remove articles, conjunctions, personal pronouns, and words that appear fewer than five times in all prospectuses.

Panel B displays statistics for IPO characteristics. There is substantial variation in the offering characteristics of firms in our sample. The mean IPO files an offer amount of approximately $195 million but issues $116 million (although the medians are very similar and near $57 million). The mean age of the firm at the time of the offering is thirteen years, but the median is significantly smaller at seven years of age. Forty-seven percent of the IPOs are classified as tech firms as defined in Loughran and Ritter (2004), while 49% have VC backing.

The average market share of the underwriter in the year prior to the offer is 2.8%, with an affiliated law firm market share of 2.3%. The average market share of the issuer's counsel is lower than that of the underwriter's counsel at 1.2%. The average market share of the auditing firm is 16.1%, which reflects the higher concentration in the auditing industry. Consistent with Lowry and Schwert (2002), IPOs are brought to market when prior returns are high, with an average return in the thirty days prior to filing of approximately 5%.

Panel C presents summary statistics describing the initial prospectus allocation. We examine various sections of the prospectus in order to capture the roles that different parts of the prospectus play in the offering process. For example, conversations with practitioners suggest that the Prospectus Summary is the main tool used by underwriters to market the IPO to potential investors. The Risk Factors and Uses of Proceeds sections, as their names imply, contain information on the various risks of the firm and how the firm intends to use the proceeds of the offer. In contrast, MD&A is intended to reflect the management's assessment of the business of the firm, not only the current financial status and strategy of the firm but also the outlook for the future. Hence, information in MD&A may be more illuminating about the firm's future prospects and should have a greater impact on pricing.

The average (and median) prospectus has just under thirty-four thousand words, of which almost 6% is the Prospectus Summary, 18% is Risk Factors, less than 1% is Use of Proceeds, and 13% consists of the MD&A. Overall, these four sections, on average, comprise 38% of the entire prospectus.[18]

Panel D shows the average number of root words that populate the word vectors. The average number of words in the vector for each section follows the document allocation noted in Panel C above. The average document has a total of almost ten thousand root words, with the Risk Factor and MD&A sections having the two largest numbers of words: 1,767 and 1,173, respectively.[19] The Prospectus Summary has an average of 609 words, but the Use of Proceeds has a mean of only seventy-three words. The small size of the Use of Proceeds section is somewhat surprising given the results of Leone, Rock, and Willenborg (2007), who find that an increase in the specificity of the intended use of proceeds reduces subsequent underpricing. This finding suggests that even small sections of the prospectus can convey important information to investors.

[18] Other sections that frequently appear in the prospectus include Capitalization, Experts, Management, Dilution, Dividend Policy, Shares Eligible for Future Sale, Legal Matters, Description of Capital Stock, Underwriting, Certain Transactions or Related Party Transactions, Principal Stockholders, Principal and Selling Stockholders, Material Tax Consequences, Certain Relationships, and Description of Securities. This list is in approximate order of frequency, and there are other, less common sections not in the list. Note that the "Management" section referred to is not MD&A. This section usually describes the profiles of key managers, their resumes, and their experience.

[19] Since the number of possible unique root words is 5,803, an average number of root words for the document as a whole of almost ten thousand means that some root words appear more frequently.

3. Sources of Content

In order to assess the source of content in the prospectus, we first construct a variable that measures the degree of similarity between documents, a measure we call "document similarity," (which is explained in more detail in Appendix A).[20] This method allows us to explore whether those who draft the prospectus use content from other sources and the degree to which content may be "standardized" across prospectuses. We also examine whether the sources of content are similar across each section of the document.

Table 2 presents a series of regressions based on the document similarities of the prospectus as a whole and each of the individual sections. The dependent variable we use is the document similarity between two initial IPO prospectuses. This is a numerical variable bounded in the interval [0,1] in which a value of 1 indicates that the two documents have exactly the same distribution of words, while a value of 0 indicates that the documents are entirely different and have no words in common. There are 1,530,375 observations for each regression (fewer appear in some specifications, as some sections are missing for a small number of IPOs). To ensure that t-statistics remain unbiased given the repeated use of each document, we report t-statistics that are adjusted for clustering by IPO.[21]

The first three explanatory variables identify the commonality of IPO i's and j's lead underwriting syndicate, whether they have the same underwriter's counsel, the same issuer's counsel, or the same auditor. To account for the situation when more than one underwriter serves as lead and i and j share at least one lead underwriter, the commonality of the lead underwriter variable is set to the proportion of common lead underwriters (number of common underwriters divided by the total number of lead underwriters). This measure has the nice property of being 0 when no lead underwriters are common and 1 when all lead underwriters are common.

Next, we include two dummy variables equal to 1 if IPOs i and j both reside in the same Fama–French forty-eight industry code and are issued within ninety days to determine whether content in the prospectus is contained in prior IPOs. We include a dummy variable if both IPOs are tech firms as identified in Loughran and Ritter (2004). Finally, we include three variables that capture how different IPO i's and j's characteristics are using the log of firm age, the IPO year, and the log of filing size.

As can be seen in Panel A, the word content of two prospectuses have greater similarity when two IPOs are brought to market by the same participants. The contribution of both industry and the timing of issuance to content similarity is

[20] A similar method was originally proposed by Markov (1913/2006) to determine authorship and is one of the first examples of a Markov chain.

[21] Document similarities are computed using many pages of text (between fifty and one hundred pages is typical for a prospectus), and prospectuses often have much obvious common content across IPOs. This means that there is a very high degree of power for measuring standard content sources accurately.

Table 2
Sources of content

Row	Common lead UW	Same UW counsel	Same issuer counsel	Same auditor	Same FF-48 industry	Within same 90 days	Both tech IPOs	Absol. age diff.	Absol. year diff.	Absol. log size diff.	R^2	Obs
					Panel A: Entire document							
(1)	0.040	0.018	0.032	0.005	0.051	0.015	0.071	−0.016	−0.010	−0.006	0.503	1,530,375
	(16.02)	(16.69)	(28.49)	(9.94)	(42.98)	(15.38)	(65.27)	(−41.28)	(−35.36)	(−37.48)		
					Panel B: Prospectus Summary							
(2)	0.048	0.012	0.019	0.006	0.038	0.012	0.040	−0.008	−0.009	−0.003	0.429	1,530,375
	(25.11)	(16.35)	(23.53)	(16.93)	(46.14)	(14.53)	(49.30)	(−26.36)	(−40.05)	(−23.53)		
					Panel C: Risk Factors							
(3)	0.024	0.019	0.032	0.004	0.056	0.013	0.077	−0.015	−0.013	−0.006	0.540	1,530,375
	(9.05)	(17.11)	(30.73)	(9.28)	(47.74)	(11.54)	(78.07)	(−41.39)	(−35.09)	(−39.46)		
					Panel D: Use of Proceeds							
(4)	0.007	0.023	0.059	0.010	0.032	0.009	0.068	−0.018	−0.016	−0.010	0.413	1,530,375
	(1.68)	(14.81)	(37.67)	(15.10)	(32.23)	(7.28)	(47.47)	(−23.52)	(−37.74)	(−44.27)		
					Panel E: MD&A							
(5)	0.069	0.019	0.036	0.006	0.037	0.016	0.050	−0.014	−0.010	−0.005	0.421	1,530,375
	(27.98)	(18.74)	(27.99)	(12.11)	(31.60)	(16.51)	(43.13)	(−29.46)	(−42.59)	(−33.14)		

OLS regressions in which the dependent variable is the Document similarity of two initial IPO prospectuses. One observation is one pair of IPOs i and j and included are all unique possible IPO pairs as observations (excluding pairs in which $i = j$). For the sample of one thousand seven hundred fifty IPOs, $\frac{1750^2-1750}{2}$ unique pairs exist, and hence, 1,530,375 observations appear in any regression. Document similarity is the dot product of the normalized (by vector length) vectors for documents i and j (this widely used method in text analysis is known as cosine similarity). (For additional information on how document similarity is measured, see Appendix A.) The independent variables measure how similar the characteristics of IPOs i and j are. The first four variables identify whether the document sections of IPOs i and j are likely written by common lead underwriters (Common lead UW), the same underwriter's counsel (Same UW counsel), the same issuer's counsel (Same issuer counsel), and the same auditor (Same auditor). When more than one lead underwriter exists, Common lead UW is set to the proportion of common lead underwriters (number of common underwriters divided by the total number of lead underwriters). The next dummy variable is 1 if IPOs i and j reside in the same Fama–French forty-eight industry code. Within same 90 days is a dummy variable identifying whether IPOs i and j are issued within ninety days of each other. Both tech IPOs is a dummy indicating whether both are in tech oriented as identified in Loughran and Ritter (2004). Finally, included are four variables measuring how different IPO i's and j's characteristics are. Each is equal to the absolute value of difference in characteristics for IPOs i and j, for each of the following: log firm age, IPO year, and log filing size. To ensure that t-statistics remain unbiased given the repeated use of each document, we include IPO fixed effects, and t-statistics are adjusted for clustering by IPO. Each regression is run for the document as a whole and for each of the four sections.

very strong, as the prospectuses are closer in content when the two IPOs are in the same industry and are offered within ninety days of each other. The results for the technology dummy and the difference in year of issuance variable also support the conclusion that concurrent offerings and common industry membership contribute much to document similarity. The table also shows that word content is less similar if the two IPO prospectuses differ in age or in size.

Once the document is parsed into the relevant sections, the effect of industry and timing of issuance remains strong. The underwriter also continues to have a large influence on each of the sections, with the exception of the Uses of Proceeds section. We can estimate the economic impact of these variables on document similarities using standard deviation units. The standard deviation of document similarities based on the whole document is 0.109 (sections have similar standard deviations). The economic magnitude of the 0.04 coefficient in Row 1 of table 2 regarding the common lead underwriter variable is 36.7% of 1 SD. Hence, IPOs having the same lead underwriter have overall document similarities that are 36.7% of 1 SD higher than those with entirely different lead underwriting syndicates. Similarly, IPOs in the same Fama–French forty-eight industry code have similarities that are 46.8% of 1 SD higher. Given that these magnitudes are large and the observation count is high, statistical significance levels are also high.

Overall, the findings of table 2 indicate that the prospectus contains standardized content from the underwriter's and the lawyers' prior prospectuses and, more importantly, from the prospectuses of other IPOs that are in recent or similar industries. Note that a limitation of document similarity is that it cannot separate content into separate standard and informative types, which would allow for deeper analysis of the role played by each. We examine this in further detail in the next section.

4. Standard Versus Informative Content

Our goal is to understand the process by which information is gathered during the IPO process. In order to assess whether a given IPO has more content related to information produced in the premarket or information gathered from investors during bookbuilding, we decompose a given prospectus word vector, $norm_{tot,i}$, into its standard and informative components. We define content that is driven by information in prior IPOs as "standard" and content that is unique to a given IPO as "informative."[22]

4.1 Method to decompose prospectus text

We propose that standard content has two primary components: content from recent or concurrent IPOs (for example, content related to recent IPO

[22] In another context and using a different method, Nelson and Pritchard (2008) try to assess how much cautionary language is standardized from one year to the next.

conditions) and content from IPOs in the same industry (industry-specific content). Recent IPOs are those that were filed in the ninety-day period preceding the current IPO's initial filing date. When computing industry content, we consider only same-industry IPOs that were filed before this ninety-day window to ensure that the two content types do not overlap. As was shown in table 2, these components have a strong relation to the wording of the prospectus, and this construction allows for enough data points to provide an estimation of standard content.[23]

We estimate IPO i's exposure to the content of recent IPOs by considering an IPO i that has K IPOs filed in the ninety-day period preceding its initial filing, whose word vectors are denoted by $words_{tot,k}$. We normalize this vector by dividing by the sum of its elements to get $norm_{tot,k}$. We define the average of the normalized vectors of recent IPOs ($norm_{rec,i}$) as

$$norm_{rec,i} = \frac{1}{K} \sum_{k=1}^{K} norm_{tot,k}.$$

Similarly, we estimate IPO i's exposure to the content of past industry IPOs by considering an IPO i that had P IPOs filed in the same Fama–French forty-eight industry code at least ninety-one days but not more than one year prior to its filing date. Defining each IPO's normalized word vector as $norm_{tot,p}$, we define the average of the normalized vectors of past industry IPOs ($norm_{ind,i}$) as

$$norm_{ind,i} = \frac{1}{P} \sum_{p=1}^{P} norm_{tot,p}.$$

We then run the following first-stage regression (without an intercept)[24] for each IPO in which one observation is one word, for a total of 5,803 observations:

$$norm_{tot,i} = a_{rec,i}\, norm_{rec,i} + a_{ind,i}\, norm_{ind,i} + \epsilon, \tag{1}$$

and define a single "standard content" variable as follows:

$$a_{standard,i} = a_{rec,i} + a_{ind,i}. \tag{2}$$

This regression compares the relative word distribution associated with IPO i with the word distribution of recent and past industry IPOs. Therefore, $a_{standard,i}$ measures the relative loading of standard content, and its interpretation is the proportion of standard words in IPO i's prospectus. We define

[23] Table 2 also indicates a strong underwriter, legal, and auditor component of standard content. Since including underwriter content reduces the sample size, we examine the influence of unique underwriter content separately in Section 5. Note that at least some underwriter (and other participant) content may be contained in either recent or past industry IPOs.

[24] Our analysis is robust to including an intercept. We have chosen to exclude the intercept, because it represents a uniform word distribution and we believe that this is not a realistic source of content.

content not explained by these two sources, the vector of the absolute value of the residuals, to be "informative content."

It is natural to think that standard and informative content might be highly correlated, as more standard content might mechanistically require less informative content. This is not the case. The raw residual in Equation (1) reflects the fact that an IPO prospectus may use hundreds of words more than the standard content vector and hundreds of other words less. Although the sum of this raw residual and the standard content coefficient must equal 1, we define informative content as the sum of the absolute values of this residual, ensuring that there is no mechanistic relationship. Indeed, our measures of standard and informative content are correlated by less than 4%, confirming that more standard content does not necessarily imply less informative content.

Note that informative content represents the absolute value of unexplained deviations from the regression model of standard content and that, therefore, standard and informative content are not simply mirrors of one another. One could think of examples of words, such as "risk" and "liability," in which fewer than expected instances would be informative to an investor. Of course, it is not necessary that such deviations be associated with words with possible negative connotations.

Standard and informative content are also estimated individually for each section in the same fashion as the document as a whole but using only the text from that particular section. For example, the estimation of the standard content for a Prospectus Summary section is based on only the word vectors from the current Prospectus Summary and the series of Prospectus Summary sections from recent and past industry IPOs.

Because each regression for each IPO has 5,803 observations, the first-stage regression in Equation (1) has ample power to fit these coefficients.[25] Because all first-stage coefficients are observed once per IPO at the time of initial filing, they can be used to determine the price impact of standard and informative content throughout the IPO process.

Table 3 Panels A and B present summary statistics for our estimation of document content. The average document in our sample has a standard content coefficient near 1. This occurs because the average document's standard content is measured using the average of both recent and past industry IPOs. When considering the full sample, the average prospectus, therefore, is similar to the average past prospectus. There is substantial variation around this loading, however, which is the subject of this analysis. The source of standard content is slightly tilted toward recent IPOs.

[25] Our results are robust to a number of alternative specifications of standard content: (i) estimation of standard content including all past IPOs instead of just recent and industry IPOs, (ii) defining past recent and industry IPOs based on closing dates rather than filing dates, (iii) expanding the definition of past industry IPOs to all observed past industry IPOs (not just those in the past year), and (iv) using a fixed number of IPOs in the estimation of recent and past industry content.

Table 3
Summary statistics on standard and informative content

Content type	Statistic	Total document	Prospectus Summary	Risk Factors	Use of Proceeds	MD&A
		Panel A: Standard vs. informative content				
Standard	Mean	1.011	1.033	1.014	1.082	1.010
	Std.	0.057	0.142	0.085	0.242	0.160
Informative	Mean	0.657	1.153	0.820	1.161	0.993
	Std.	0.075	0.105	0.116	0.214	0.135
		Panel B: Recent and past industry content				
Recent	Mean	0.579	0.661	0.639	0.837	0.662
	Std.	0.441	0.417	0.503	0.669	0.502
Past industry	Mean	0.432	0.372	0.376	0.245	0.348
	Std.	0.434	0.401	0.498	0.662	0.484

Summary statistics on the coefficients from the first-stage regressions on document content for one thousand seven hundred fifty IPOs issued in the United States from August 1996 to October 2005, excluding firms with an issue price less than $5, ADRs, financial firms, unit IPOs, dual class IPOs, and REITs. In Panels A and B, the coefficients on Standard, Informative, Recent ($a_{ind,i}$), and Past industry ($a_{ind,i}$) content are from the first-stage regression for each IPO i: $norm_{tot,i} = a_{rec,i}\ norm_{rec,i} + a_{ind,i}\ norm_{ind,i} + \epsilon$, where Standard content is the sum of the coefficients $a_{rec,i}$ and $a_{ind,i}$, and Informative content is the sum of the absolute residuals.

In order to assess the determinants of the level of standard and informative content, table 4 uses, as the dependent variable, the estimates of standard and informative content from the regression defined in Equation (1) above. Consistent with our interpretation of Khanna, Noe, and Sonti (2008), the variable *Log Number of UW IPO Ratio* has a large impact on the level of standard and informative content. This variable is defined as the natural logarithm of 1 plus the total number of IPOs filed by all lead underwriters in the given IPO's syndicate in the year prior to the current IPO's filing (year $t-1$), divided by the number of IPOs filed by the same underwriters in years $t-2$ and $t-3$. By scaling recent underwriter activity by past activity, this variable is a proxy for the amount of unexpected demand on the underwriter's resources and employees. Our results suggest that when there are many IPOs competing for the underwriter's attention, they may not have the necessary resources to invest in enhanced due diligence and will, therefore, generate more standard and less informative content.

Other firm and offering characteristics, such as underwriter market share and offer size, are related to the amount of informative content for only a few of the sections. However, VC-backed IPOs have lower informative content and greater standard content. One interpretation is that VC-backed IPOs may be more likely to have proprietary information and that, therefore, investing in information production may not be as worthwhile if its subsequent disclosure would reveal valuable information to rivals.

4.2 Relation to pricing

We suggest that the relative amounts of standard and informative content in an IPO's prospectus proxy for the effort expended by issuers and underwriters to

Table 4
Determinants of standard and informative content

Row	Section	Log firm age	UW $ mkt share	UW law $ mkt share	Iss law $ mkt share	Auditor $ mkt share	VC dummy	Pre-file Nasdaq return	Log # 90 day IPOs	Log # industry IPOs	Log dollars filed	Log # UW IPO ratio	Year+Ind fixed effects	R^2
						Panel A: Standard content								
(1)	Whole document	−0.000	−0.169	−0.048	0.078	0.028	0.006	−0.001	0.000	0.003	−0.003	0.006	Yes	0.109
		(−0.24)	(−2.58)	(−1.51)	(1.00)	(1.54)	(1.77)	(−0.09)	(0.08)	(0.46)	(−1.65)	(2.52)		
(2)	Prospectus Summary	0.006	−0.211	−0.177	0.116	0.120	0.029	0.016	−0.005	0.006	−0.007	0.006	Yes	0.100
		(1.54)	(−1.29)	(−1.23)	(0.83)	(2.71)	(2.29)	(0.50)	(−0.46)	(0.52)	(−1.83)	(1.14)		
(3)	Risk Factors	0.007	−0.308	0.046	0.018	0.011	0.019	−0.007	−0.000	0.012	−0.009	0.002	Yes	0.131
		(2.83)	(−2.64)	(0.62)	(0.21)	(0.48)	(4.71)	(−0.34)	(−0.02)	(1.71)	(−2.08)	(0.91)		
(4)	Use of Proceeds	−0.012	0.643	0.186	0.391	0.265	0.074	0.154	0.003	−0.014	0.015	0.009	Yes	0.106
		(−1.78)	(2.44)	(0.93)	(1.10)	(3.31)	(4.16)	(2.32)	(0.25)	(−1.08)	(1.86)	(0.72)		
(5)	MD&A	0.018	−0.265	−0.128	−0.072	0.030	0.027	0.056	0.015	−0.012	−0.001	0.006	Yes	0.079
		(3.34)	(−1.53)	(−1.14)	(−0.42)	(0.67)	(2.89)	(1.42)	(1.57)	(−0.59)	(−0.21)	(1.01)		
						Panel B: Informative content								
(6)	Whole document	−0.004	−0.072	−0.097	−0.212	−0.003	−0.043	−0.041	−0.009	−0.022	−0.001	−0.006	Yes	0.460
		(−2.00)	(−1.32)	(−1.91)	(−4.46)	(−0.12)	(−11.45)	(−2.03)	(−1.79)	(−3.75)	(−0.53)	(−2.64)		
(7)	Prospectus Summary	−0.008	−0.209	−0.214	−0.101	−0.007	0.008	−0.047	−0.011	−0.019	−0.020	0.000	Yes	0.293
		(−3.02)	(−1.48)	(−3.73)	(−1.07)	(−0.22)	(1.92)	(−1.72)	(−1.00)	(−1.93)	(−7.75)	(0.05)		
(8)	Risk Factors	0.013	0.210	−0.161	−0.283	0.010	−0.054	−0.008	−0.006	−0.027	0.004	−0.006	Yes	0.395
		(3.85)	(2.18)	(−2.21)	(−2.74)	(0.28)	(−8.99)	(−0.24)	(−0.83)	(−3.19)	(1.23)	(−1.35)		
(9)	Use of Proceeds	0.010	1.040	0.092	−0.055	0.064	−0.065	0.025	−0.025	−0.047	0.026	0.000	Yes	0.259
		(1.91)	(3.88)	(0.58)	(−0.24)	(1.17)	(−5.07)	(0.67)	(−1.53)	(−3.06)	(4.39)	(0.01)		
(10)	MD&A	−0.015	−0.219	−0.146	−0.229	−0.041	−0.063	−0.033	−0.007	−0.020	−0.011	−0.011	Yes	0.359
		(−5.04)	(−1.81)	(−1.61)	(−2.56)	(−1.28)	(−9.86)	(−1.07)	(−0.80)	(−1.65)	(−3.40)	(−2.70)		

OLS regressions with yearly fixed effects are presented for one thousand seven hundred fifty IPOs issued in the United States from August 1996 to October 2005, excluding firms with an issue price less than $5, ADRs, financial firms, unit IPOs, dual class IPOs, and REITs. The dependent variable is either the Standard content (Panel A) or Informative content (Panel B) of the total document or a document section based on the following first-stage regression for each IPO i: $norm_{tot,i} = a_{rec,i}\, norm_{rec,i} + a_{ind,i}\, norm_{ind,i} + \epsilon$. Standard content is the sum of the coefficients $a_{rec,i}$ and $a_{ind,i}$, and Informative content is the sum of the absolute residuals. The independent variables are as follows: Log firm age is the log of the IPO year minus the firm's founding date, where founding dates are obtained from the Field–Ritter dataset, as used in Field and Karpoff (2002). UW $ mkt share is the lead underwriter's dollar market share in the past calendar year. UW law $ mkt share is the underwriting firm's legal counsel's dollar market share in the past calendar year. Issuer law $ mkt share is the issuer firm's legal counsel's dollar market share in the past calendar year. The VC dummy is equal to 1 if a firm is VC financed. Auditor $ mkt share is the auditor's dollar market share in the past calendar year. Pre-file Nasdaq return is the NASDAQ return for the thirty trading days preceding the filing date. The Log # 90 day IPOs and Log # industry IPOs are the natural logarithms of the number of IPOs used to construct the past ninety days and past industry components of standard content, respectively. Log # UW IPO ratio is the natural logarithm of one plus the total number of IPOs filed by all lead underwriters in the given IPO's syndicate in the year prior to the current IPO's filing (year $t-1$), divided by the number of IPOs filed by the same underwriters in years $t-2$ and $t-3$. The Log dollars filed is the log of the original filing amount. Year+Ind fixed effects are also included, where industry definitions are based on the Fama–French forty-eight industry codes. t-Statistics are in parentheses and are adjusted for clustering within industries and years.

gather information in the premarket. In particular, we argue that the greater the amount of information production in the premarket by the underwriter and issuer, the larger will be the amount of informative content in the initial prospectus. This should lead to greater pricing accuracy in the expected offer price and lower revelation of information from investors during bookbuilding. Thus, we hypothesize that the larger the informative content, the smaller will be the price change, both absolute and raw, during bookbuilding, and the lower will be the subsequent underpricing.

We predict the opposite for standard content. Since standard content is information that is contained in recent and past industry IPOs, a large amount of standard content is consistent with lower information production in the premarket. We hypothesize that the larger the standard content, the more likely it will be that information is gathered from investors during bookbuilding. The larger the standard content, the greater will be the price change, both absolute and raw, and underpricing during bookbuilding.

We examine the impact of both standard and informative content on three pricing variables: the change in offer price from the midpoint of the initial file range (ΔP), the absolute value of the change in offer price (abs(ΔP)), and the initial return (IR). In order to control for any potential relationships between firm and offering characteristics in the choice between standard and informative content, we hold constant a number of firm-specific variables as well as include year and industry effects. All regressions have standard errors that are adjusted for clustering by year and industry.

Table 5 presents a series of regressions for the document as a whole as well as the four sections on the relation of standard versus informative content to IPO pricing. Almost uniformly, the greater the standard content of the prospectus, the greater is ΔP or absolute ΔP and subsequent underpricing. These findings not only support our central predictions regarding pricing accuracy but also point to potential lowballing of the initial offer price when there is less premarket due diligence. The opposite holds true for informative content. The greater the informative content, the smaller is ΔP or absolute ΔP and initial return. The only exception is the Prospectus Summary.

To ensure that our results are robust to the exclusion of tech firms, table 6 reproduces table 5 for the 927 IPOs that exclude technology firms as identified by Loughran and Ritter (2004). The results remain relatively similar.

The economic magnitude of the relation between informative and standard content for the sample as a whole is presented in table 7. A 1 SD change in standard content in the entire document, for example, is associated with an approximately 1.5% increase in the level of the offer price from the filing midpoint and a 4% increase in initial returns. Although significant for many sections, the absolute change in the offer price is economically insignificant for the document as a whole. A 1 SD change in informative content, on the other hand, is associated with a reduction in the level and absolute value of the change in offer price from the filing midpoint by around 2%. Initial returns de-

Table 5
Relation of standard and informative content to price adjustments and initial returns

Row	Section	Dependent variable	Standard content	Informative content	Log firm age	UW $ market share	UW law $ market share	Iss law $ market share	Auditor $ market share	VC dummy	Pre-file Nasdaq return	Log dollars filed	Year+Ind fixed effects	R^2
(1)	TOTAL	ΔP	0.261	−0.304	−0.014	2.402	−0.213	0.409	0.049	0.018	0.111	−0.017	Yes	0.207
			(2.36)	(−2.09)	(−2.18)	(6.56)	(−0.73)	(1.030)	(0.72)	(1.18)	(1.33)	(−1.67)		
(2)	TOTAL	abs(ΔP)	0.052	−0.391	−0.019	0.548	−0.195	0.129	0.009	0.033	0.157	0.010	Yes	0.165
			(0.64)	(−3.86)	(−2.70)	(1.54)	(−1.03)	(0.50)	(0.20)	(3.67)	(2.13)	(1.38)		
(3)	TOTAL	IR	0.722	−1.077	−0.060	5.260	−0.006	1.696	−0.273	0.121	0.385	−0.040	Yes	0.256
			(3.16)	(−3.75)	(−2.86)	(2.63)	(−0.01)	(1.36)	(−1.51)	(3.91)	(2.09)	(−2.33)		
(4)	PS	ΔP	0.142	0.060	−0.014	2.422	−0.158	0.484	0.040	0.029	0.119	−0.015	Yes	0.207
			(1.98)	(0.57)	(−2.21)	(6.67)	(−0.58)	(1.55)	(0.60)	(1.78)	(1.35)	(−1.32)		
(5)	PS	abs(ΔP)	0.065	0.089	−0.018	0.599	−0.129	0.217	0.004	0.048	0.171	0.013	Yes	0.158
			(1.50)	(1.07)	(−2.61)	(1.69)	(−0.70)	(0.86)	(0.10)	(5.81)	(2.19)	(1.63)		
(6)	PS	IR	0.382	0.112	−0.059	5.318	0.154	1.949	−0.295	0.161	0.413	−0.035	Yes	0.252
			(3.36)	(0.61)	(−2.93)	(2.66)	(0.27)	(1.55)	(−1.64)	(4.41)	(2.17)	(−2.27)		
(7)	RF	ΔP	0.029	−0.214	−0.011	2.433	−0.232	0.434	0.059	0.021	0.119	−0.016	Yes	0.206
			(0.44)	(−2.84)	(−1.72)	(6.55)	(−0.79)	(1.46)	(0.86)	(1.22)	(1.42)	(−1.62)		
(8)	RF	abs(ΔP)	0.136	−0.270	−0.015	0.665	−0.210	0.137	0.013	0.033	0.170	0.012	Yes	0.169
			(2.43)	(−5.80)	(−2.26)	(1.87)	(−1.13)	(0.54)	(0.28)	(3.52)	(2.29)	(1.73)		
(9)	RF	IR	0.290	−0.613	−0.051	5.432	−0.050	1.804	−0.247	0.133	0.418	−0.036	Yes	0.253
			(1.42)	(−3.54)	(−2.89)	(2.70)	(−0.08)	(1.47)	(−1.39)	(3.53)	(2.29)	(−2.01)		

(continued)

Table 5
Continued

Row	Section	Dependent variable	Standard content	Informative content	Log firm age	UW $ market share	UW law $ market share	Iss law $ market share	Auditor $ market share	VC dummy	Pre-file Nasdaq return	Log dollars filed	Year+Ind fixed effects	R^2
(10)	USE	ΔP	0.104	−0.076	−0.011	2.393	−0.207	0.450	0.034	0.021	0.107	−0.017	Yes	0.206
			(3.61)	(−1.73)	(−1.81)	(6.10)	(−0.72)	(1.54)	(0.51)	(1.29)	(1.28)	(−1.56)		
(11)	USE	abs(ΔP)	0.113	−0.111	−0.015	0.610	−0.169	0.166	−0.011	0.034	0.158	0.011	Yes	0.167
			(6.65)	(−3.60)	(−2.19)	(1.60)	(−0.91)	(0.70)	(−0.24)	(4.08)	(2.14)	(1.52)		
(12)	USE	IR	0.396	−0.397	−0.048	5.375	0.031	1.807	−0.329	0.117	0.376	−0.036	Yes	0.260
			(6.27)	(−3.38)	(−2.44)	(2.52)	(0.05)	(1.55)	(−1.87)	(3.57)	(2.16)	(−2.03)		
(13)	MDA	ΔP	0.123	−0.336	−0.021	2.340	−0.228	0.426	0.039	0.008	0.106	−0.021	Yes	0.215
			(2.47)	(−4.10)	(−3.22)	(6.33)	(−0.82)	(1.37)	(0.58)	(0.52)	(1.30)	(−1.98)		
(14)	MDA	abs(ΔP)	0.059	−0.194	−0.022	0.541	−0.179	0.175	0.002	0.036	0.162	0.008	Yes	0.162
			(1.63)	(−2.89)	(−3.19)	(1.52)	(−0.95)	(0.69)	(0.04)	(3.99)	(2.18)	(1.09)		
(15)	MDA	IR	0.378	−0.906	−0.077	5.121	−0.016	1.800	−0.300	0.104	0.379	−0.050	Yes	0.263
			(3.13)	(−5.26)	(−3.82)	(2.54)	(−0.03)	(1.43)	(−1.70)	(3.20)	(2.10)	(−2.97)		

OLS regressions with yearly fixed effects are presented for one thousand seven hundred fifty IPOs issued in the United States from August 1996 to October 2005, excluding firms with an issue price less than $5, ADRs, financial firms, unit IPOs, dual class IPOs, and REITs. The dependent variable is either the Change in price from the filing date midpoint to the IPO offer price (ΔP), the Absolute value of the change in offer price (abs(ΔP)), or the Initial return (IR). We report the results for the prospectus as a whole (TOTAL) and all four sections: Prospectus Summary (PS), Risk Factors (RF), Use of Proceeds (USE), and MD&A. Two key independent variables are the IPO's standard and informative content, which are based on the following first-stage regression for each IPO i: $norm_{tot,i} = a_{rec,i}\, norm_{rec,i} + a_{ind,i}\, norm_{ind,i} + \epsilon$. Standard content is the sum of the coefficients $a_{rec,i}$ and $a_{ind,i}$, and Informative content is the sum of the absolute residuals. The remaining independent variables are as follows: Log firm age is the log of the IPO year minus the firm's founding date, where founding dates are obtained from the Field–Ritter dataset, as used in Field and Karpoff (2002) and Loughran and Ritter (2004). UW $ market share is the lead underwriter's dollar market share in the past calendar year. UW law $ market share is the underwriting firm's legal counsel's dollar market share in the past calendar year. Issuer law $ market share is the issuer firm's legal counsel's dollar market share in the past calendar year. The VC dummy is equal to 1 if a firm is VC financed. Auditor $ market share is the auditor's dollar market share in the past calendar year. Pre-file Nasdaq return is the NASDAQ return for the thirty trading days preceding the filing date. The Log dollars filed is the log of the original filing amount. Year+Ind fixed effects are also included, where industry definitions are based on the Fama–French forty-eight industry code. The Tech dummy, based on Loughran and Ritter (2004), is also included but not shown to conserve space. t-Statistics are in parentheses and are adjusted for clustering within industries and years.

Table 6
Relation of standard and informative content to price adjustments and initial returns (excluding technology IPOs)

Row	Section	Dependent variable	Standard content	Informative content	Log firm age	UW $ market share	UW law $ market share	Iss law $ market share	Auditor $ market share	VC dummy	Pre-file Nasdaq return	Log dollars filed	Year+Ind fixed effects	R^2
(1)	TOTAL	ΔP	0.307	−0.170	0.002	2.471	−0.130	−0.075	−0.106	0.005	0.011	−0.021	Yes	0.188
			(2.24)	(−1.00)	(0.31)	(4.08)	(−0.46)	(−0.24)	(−1.49)	(0.24)	(0.14)	(−1.99)		
(2)	TOTAL	abs(ΔP)	−0.006	−0.417	−0.014	0.384	0.005	−0.390	0.039	0.030	0.102	−0.002	Yes	0.166
			(−0.08)	(−3.75)	(−2.66)	(0.81)	(0.04)	(−1.96)	(0.81)	(2.43)	(1.08)	(−0.31)		
(3)	TOTAL	IR	0.813	−0.719	−0.025	3.175	0.262	−0.184	−0.348	0.101	0.488	−0.030	Yes	0.232
			(3.05)	(−2.90)	(−1.94)	(2.36)	(0.51)	(−0.21)	(−2.02)	(2.34)	(2.62)	(−1.64)		
(4)	PS	ΔP	0.155	−0.022	0.001	2.454	−0.103	−0.007	−0.121	0.008	0.011	−0.021	Yes	0.187
			(1.59)	(−0.21)	(0.13)	(4.19)	(−0.38)	(−0.02)	(−1.67)	(0.46)	(0.13)	(−1.97)		
(5)	PS	abs(ΔP)	0.089	−0.010	−0.013	0.416	0.024	−0.294	0.028	0.044	0.103	−0.001	Yes	0.151
			(1.48)	(−0.14)	(−2.25)	(0.90)	(0.18)	(−1.45)	(0.56)	(3.88)	(1.06)	(−0.13)		
(6)	PS	IR	0.443	0.034	−0.027	3.182	0.358	0.093	−0.387	0.120	0.489	−0.027	Yes	0.230
			(3.01)	(0.25)	(−1.88)	(2.46)	(0.72)	(0.10)	(−2.30)	(2.66)	(2.59)	(−1.54)		
(7)	RF	ΔP	0.030	−0.200	0.004	2.524	−0.147	−0.017	−0.092	−0.000	0.007	−0.020	Yes	0.187
			(0.38)	(−1.87)	(0.62)	(4.01)	(−0.53)	(−0.05)	(−1.32)	(−0.00)	(0.09)	(−1.92)		
(8)	RF	abs(ΔP)	0.135	−0.281	−0.010	0.559	−0.031	−0.339	0.041	0.028	0.105	0.002	Yes	0.173
			(2.61)	(−4.36)	(−2.24)	(1.15)	(−0.22)	(−1.60)	(0.86)	(2.31)	(1.10)	(0.32)		
(9)	RF	IR	0.272	−0.528	−0.020	3.387	0.184	0.018	−0.315	0.096	0.481	−0.023	Yes	0.228
			(1.84)	(−2.89)	(−1.66)	(2.42)	(0.37)	(0.02)	(−1.96)	(2.16)	(2.56)	(−1.36)		

(continued)

Table 6
Continued

Row	Section	Dependent variable	Standard content	Informative content	Log firm age	UW $ market share	UW law $ market share	Iss law $ market share	Auditor $ market share	VC dummy	Pre-file Nasdaq return	Log dollars filed	Year+Ind fixed effects	R^2
(10)	USE	ΔP	0.083	−0.070	0.003	2.494	−0.147	0.014	−0.106	0.001	−0.001	−0.021	Yes	0.185
			(1.92)	(−1.52)	(0.47)	(4.03)	(−0.54)	(0.04)	(−1.53)	(0.04)	(−0.02)	(−1.96)		
(11)	USE	abs(ΔP)	0.121	−0.127	−0.011	0.521	−0.001	−0.282	0.026	0.028	0.091	0.002	Yes	0.169
			(3.87)	(−3.39)	(−2.12)	(1.05)	(−0.01)	(−1.37)	(0.53)	(2.35)	(0.98)	(0.21)		
(12)	USE	IR	0.295	−0.292	−0.020	3.354	0.229	0.124	−0.354	0.087	0.449	−0.024	Yes	0.232
			(3.98)	(−2.92)	(−1.55)	(2.38)	(0.47)	(0.13)	(−2.17)	(2.16)	(2.51)	(−1.28)		
(13)	MDA	ΔP	0.131	−0.234	−0.003	2.420	−0.134	−0.007	−0.106	−0.002	−0.004	−0.026	Yes	0.191
			(2.84)	(−2.40)	(−0.46)	(4.01)	(−0.48)	(−0.02)	(−1.54)	(−0.12)	(−0.05)	(−2.20)		
(14)	MDA	abs(ΔP)	0.063	−0.208	−0.016	0.390	−0.003	−0.317	0.034	0.035	0.098	−0.004	Yes	0.160
			(1.87)	(−2.45)	(−2.66)	(0.84)	(−0.02)	(−1.62)	(0.71)	(3.03)	(1.04)	(−0.55)		
(15)	MDA	IR	0.358	−0.658	−0.038	3.052	0.250	0.060	−0.345	0.092	0.446	−0.040	Yes	0.236
			(3.82)	(−3.91)	(−2.63)	(2.32)	(0.50)	(0.07)	(−2.10)	(2.18)	(2.40)	(−2.16)		

OLS regressions with yearly fixed effects are presented for 927 IPOs issued in the United States from August 1996 to October 2005, excluding technology firms (as identified by Loughran and Ritter (2004), firms with an issue price less than $5, ADRs, financial firms, unit IPOs, dual class IPOs, and REITs. The dependent variable is either the Change in price from the filing date midpoint to the IPO offer price (ΔP), the Absolute value of the change in offer price (abs(ΔP)), or the Initial return (IR). We report results for the prospectus as a whole (TOTAL) and all four sections: Prospectus Summary (PS), Risk Factors (RF), Use of Proceeds (USE), and MD&A. Two key independent variables are the IPO's standard and informative content, which are based on the following first-stage regression for each IPO i: $norm_{tot,i} = a_{rec,i}\, norm_{rec,i} + a_{ind,i}\, norm_{ind,i} + \epsilon$. Standard content is the sum of the coefficients $a_{rec,i}$ and $a_{ind,i}$, and Informative content is the sum of the absolute residuals. The remaining independent variables are as follows. Log firm age is the log of the IPO year minus the firm's founding date, where founding dates are obtained from the Field–Ritter dataset, as used in Field and Karpoff (2002) and Loughran and Ritter (2004). UW $ market share is the lead underwriter's dollar market share in the past calendar year. UW law $ market share is the underwriting firm's legal counsel's dollar market share in the past calendar year. Issuer law $ market share is the issuer firm's legal counsel's dollar market share in the past calendar year. The VC dummy is equal to 1 if a firm is VC financed. Auditor $ market share is the auditor's dollar market share in the past calendar year. Pre-file Nasdaq return is the NASDAQ return for the thirty trading days preceding the filing date. The Log dollars filed is the log of the original filing amount. Year+Ind fixed effects are also included, where industry definitions are based on the Fama–French forty-eight industry codes. t-Statistics are in parentheses and are adjusted for clustering within industries and years.

Table 7
Economic significance of standard and informative content

Section	Content	Standard deviation	Coefficient	Economic significance	*t*-Stat	Section	Content	Standard deviation	Coefficient	Economic significance	*t*-Stat
				Panel A: Dependent variable: ΔP							
TOTAL	Standard	0.057	0.261	1.49%	2.36	TOTAL	Informative	0.075	−0.304	−2.28%	−2.09
PS	Standard	0.142	0.142	2.02%	1.98	PS	Informative	0.105	0.060	0.63%	0.57
RF	Standard	0.085	0.029	0.25%	0.44	RF	Informative	0.116	−0.214	−2.48%	−2.84
USE	Standard	0.242	0.104	2.52%	3.61	USE	Informative	0.214	−0.076	−1.63%	−1.73
MD&A	Standard	0.160	0.123	1.97%	2.47	MDA	Informative	0.135	−0.336	−4.54%	−4.10
				Panel B: Dependent variable: abs(ΔP*)*							
TOTAL	Standard	0.057	0.052	0.30%	0.64	TOTAL	Informative	0.075	−0.391	−2.93%	−3.86
PS	Standard	0.142	0.065	0.92%	1.50	PS	Informative	0.105	0.089	0.93%	1.07
RF	Standard	0.085	0.136	1.16%	2.43	RF	Informative	0.116	−0.270	−3.13%	−5.80
USE	Standard	0.242	0.113	2.73%	6.65	USE	Informative	0.214	−0.111	−2.38%	−3.60
MDA	Standard	0.160	0.059	0.94%	1.63	MDA	Informative	0.135	−0.194	−2.62%	−2.89
				Panel C: Dependent variable: IR							
TOTAL	Standard	0.057	0.722	4.12%	3.16	TOTAL	Informative	0.075	−1.077	−8.08%	−3.75
PS	Standard	0.142	0.382	5.42%	3.36	PS	Informative	0.105	0.112	1.18%	0.61
RF	Standard	0.085	0.290	2.47%	1.42	RF	Informative	0.116	−0.613	−7.11%	−3.54
USE	Standard	0.242	0.396	9.58%	6.27	USE	Informative	0.214	−0.397	−8.50%	−3.38
MDA	Standard	0.160	0.378	6.05%	3.13	MDA	Informative	0.135	−0.906	−12.23%	−5.26

Economic significance of the effect of standard and informative content on pricing is reported for the regressions presented in table 5. Economic significance is defined as the coefficient times the standard deviation. We report results for the prospectus as a whole (TOTAL) and all four sections: Prospectus Summary (PS), Risk Factors (RF), Use of Proceeds (USE), and MD&A.

crease by 8%, consistent with greater pricing accuracy in the initial offer price as well as decreased compensation in the form of underpricing for information production during bookbuilding.[26]

We find little impact of informative content in the Prospectus Summary on pricing. This may be due to the role of this section as a marketing tool used by underwriters to generate interest from investors. Text uniqueness may not be as meaningful in this section as in others with more substance, such as the Risk Factors or MD&A. This is apparent when examining the impact of a change in the content of MD&A on subsequent underpricing. A 1 *SD* change in standard content equates to a 6% increase in initial returns, while a similar change in informative content equates to a 12% reduction in initial return. This finding underscores the importance of management disclosure in IPO pricing. While the IPO literature has focused primarily on the role of the underwriter, the ability of management to influence offer prices has not been studied. These results highlight the potentially important role that management and content in the MD&A section play in the offering process.

We acknowledge the difficulty in determining causality from these tests. Although we attempt to hold constant firm characteristics, offering, industry, and year effects, we do not have a natural experiment in which a random subset of firms are forced to disclose only standard content.[27] From conversations with participants who draft the prospectus, however, we understand that the amount and type of information to disclose has a discretionary component, which we conjecture is based on the trade-offs of acquiring more information in the premarket or gathering information during bookbuilding. If the decision to invest in premarket due diligence is a choice variable, then a positive (negative) association between the amount of informative (standard) content and subsequent pricing is consistent with greater information acquisition and disclosure in the premarket, reducing the need for information gathered from investors during bookbuilding.

4.3 Effect on IPO expenses

Our findings indicate that greater informative content increases offer price accuracy and reduces underpricing. The opposite is true for standard content. While we conjecture that differences in premarket due diligence are responsible for our results, there does exist an alternate explanation that is invariant to the level of effort. Under this scenario, the amount of effort expended in premarket due diligence is constant across all IPOs, but there exist differences in the amount of information that is disclosed or withheld in the initial

[26] Our results are in line with Leone, Rock, and Willenborg (2007), who find that a 1 *SD* increase in the specificity (informativeness) of the uses of proceeds results in an 11% reduction in underpricing. We find an 8.5% reduction in underpricing for a 1 *SD* increase in the informativeness of the Use of Proceeds section. An advantage of our method is that it does not require any hand collection of data.

[27] We also cannot conceive of an instrumental variable that captures the choice of standard versus informative content but does not affect pricing.

prospectus. (Note that it may not be possible to completely distinguish between these two stories because the trade-off to engage in premarket due diligence may depend on whether the firm intends to disclose valuable information.)

However, we can shed additional light on whether differences exist in the amount of premarket effort by examining the relationship between content type and issuer expenses. The intuition is that higher expenses are directly related to higher effort levels in the premarket but are not significantly affected by the decision to withhold or disclose known information. Higher premarket information gathering should generate larger fees to compensate lawyers, accountants, and investment bankers for their additional time devoted to enhanced due diligence. Therefore, the effort-based interpretation of the source of informative content predicts that more informative content should be associated with higher issuer expenses.

This predicted link to expenses is particularly true for lawyers and accountants, whose compensation is based on the amount of time spent on the particular transaction, rather than being a percentage of gross proceeds. The predicted effect on underwriter compensation is less clear, because spreads reflect a variety of underwriter activities and risks (Torstila 2001), many not related to due diligence.[28] It is also possible that reduced effort in the premarket may increase effort in other areas such that total underwriter fees are unaffected by the type of content.

Table 8 presents evidence on the relation between standard and informative content and issuer expenses. Both legal and accounting fees are significantly higher for documents with greater informative content. The economic impact on legal fees, in particular, is especially large. A 1 *SD* increase in the informative content of the entire document increases the average legal fee ($457,911) by almost 24%.[29] Legal fees are reduced by almost 8% for a 1 *SD* increase in standard content. However, this latter economic magnitude is not significant at conventional levels.

While accounting fees are positively related to both the informative and the standard content of the document as a whole, the economic impact for a 1 *SD* increase in informative content (42%) is almost double the impact of a 1 *SD* increase in standard content (24%).

In order to address the complexity of the gross spread, we consider its three components: management fee, underwriting fee, and selling fee. Enhanced due diligence should be reflected most directly as part of the management fee, as the lead manager is most responsible for assisting the issuer in drafting the

[28] It could also be the case that gross spreads are a substitute for underpricing. If the underwriter maximizes its total compensation, fees plus underpricing, greater informative content could result in higher gross spreads to offset the reduction in underpricing. Note that this argument, however, would not affect lawyer or auditor fees.

[29] Economic significance is calculated as follows: a 1 *SD* increase in informative content (0.075) results in a 0.094% increase (0.075*1.255) in the percentage legal fee to gross proceeds. This results in a dollar increase of (0.94%*$116 million = $109,185), which is 23.84% of the mean level of legal fees ($457,911).

Table 8
Relation of standard and informative content to IPO expenses

Row	Section	Dep. variable	Standard content	Informative content	Log firm age	UW $ market share	UW law $ market share	Iss law $ market share	Log issue size	R^2	Obs
(1)	ALL	Spread	0.626	0.577	−0.060	2.846	−0.456	−0.763	−0.593	0.516	1,748
			(2.86)	(2.57)	(−4.54)	(3.09)	(−1.53)	(−1.63)	(−11.94)		
(2)	ALL	Mgmt fee	0.266	0.637	−0.019	−0.878	−0.217	−0.506	−0.183	0.380	1,650
			(1.26)	(4.67)	(−2.83)	(−1.88)	(−1.48)	(−2.62)	(−9.37)		
(3)	ALL	Und. fee	0.051	0.291	0.002	−0.868	−0.131	−0.393	−0.204	0.412	1,648
			(0.46)	(2.24)	(0.31)	(−1.75)	(−0.92)	(−1.48)	(−12.36)		
(4)	ALL	Sell fee	0.569	0.297	−0.030	−0.608	−0.310	−1.158	−0.281	0.315	1,734
			(2.56)	(1.66)	(−2.56)	(−0.54)	(−0.89)	(−2.59)	(−9.27)		
(5)	ALL	Legal fee	−0.523	1.255	−0.015	−0.774	1.684	0.497	−0.532	0.454	1,400
			(−1.26)	(3.57)	(−0.74)	(−0.97)	(3.12)	(0.92)	(−15.60)		
(6)	ALL	Acct fee	1.344	1.842	−0.087	0.339	1.598	−0.056	−0.352	0.255	1,401
			(2.07)	(4.43)	(−2.33)	(0.30)	(3.10)	(−0.10)	(−11.68)		
(7)	PS	Spread	−0.074	0.111	−0.059	2.656	−0.540	−0.861	−0.592	0.513	1,748
			(−0.72)	(0.47)	(−4.62)	(2.91)	(−1.72)	(−1.80)	(−11.30)		
(8)	PS	Mgmt fee	0.038	0.001	−0.018	−0.976	−0.296	−0.680	−0.181	0.366	1,650
			(0.44)	(0.01)	(−2.64)	(−2.06)	(−2.00)	(−3.30)	(−9.03)		
(9)	PS	Und. fee	−0.018	−0.065	0.002	−0.918	−0.184	−0.479	−0.205	0.410	1,648
			(−0.35)	(−0.55)	(0.28)	(−1.83)	(−1.24)	(−1.86)	(−12.52)		
(10)	PS	Sell fee	0.114	−0.328	−0.034	−0.757	−0.423	−1.251	−0.289	0.314	1,734
			(0.91)	(−2.00)	(−2.97)	(−0.69)	(−1.14)	(−2.69)	(−8.93)		
(11)	PS	Legal fee	−0.268	−0.013	−0.012	−0.831	1.506	0.188	−0.532	0.446	1,400
			(−1.80)	(−0.06)	(−0.59)	(−1.08)	(2.70)	(0.34)	(−14.73)		
(12)	PS	Acct fee	0.257	−0.562	−0.094	0.029	1.221	−0.539	−0.371	0.239	1,401
			(0.83)	(−1.64)	(−2.40)	(0.03)	(2.43)	(−0.99)	(−12.09)		
(13)	RF	Spread	−0.042	0.203	−0.064	2.630	−0.511	−0.801	−0.596	0.513	1,748
			(−0.22)	(1.40)	(−4.84)	(2.90)	(−1.66)	(−1.74)	(−11.40)		
(14)	RF	Mgmt fee	−0.124	0.206	−0.021	−1.032	−0.257	−0.592	−0.186	0.370	1,650
			(−1.46)	(1.96)	(−2.82)	(−2.21)	(−1.75)	(−3.05)	(−8.82)		
(15)	RF	Und. fee	−0.003	0.028	0.002	−0.911	−0.161	−0.465	−0.204	0.410	1,648
			(−0.04)	(0.38)	(0.26)	(−1.78)	(−1.10)	(−1.78)	(−12.26)		
(16)	RF	Sell fee	0.281	0.188	−0.036	−0.728	−0.358	−1.145	−0.278	0.314	1,734
			(2.11)	(1.23)	(−3.15)	(−0.67)	(−1.02)	(−2.63)	(−8.87)		
(17)	RF	Legal fee	−0.888	0.240	−0.013	−0.902	1.588	0.233	−0.547	0.453	1,400
			(−3.81)	(1.34)	(−0.58)	(−1.15)	(2.81)	(0.41)	(−16.12)		
(18)	RF	Acct fee	−0.635	0.637	−0.097	−0.255	1.367	−0.227	−0.369	0.243	1,401
			(−2.44)	(2.55)	(−2.40)	(−0.24)	(2.72)	(−0.50)	(−12.12)		

(continued)

Table 8
Continued

(19)	USE	Spread	−0.220	−0.001	−0.064	2.796	−0.501	−0.758	−0.590	0.517	1,748
			(−2.23)	(−0.01)	(−4.96)	(3.06)	(−1.60)	(−1.74)	(−11.63)		
(20)	USE	Mgmt fee	−0.164	0.069	−0.023	−0.936	−0.276	−0.572	−0.183	0.377	1,650
			(−4.35)	(2.03)	(−3.34)	(−2.04)	(−1.89)	(−3.17)	(−9.37)		
(21)	USE	Und. fee	−0.086	0.059	−0.001	−0.891	−0.154	−0.418	−0.206	0.412	1,648
			(−2.58)	(2.21)	(−0.09)	(−1.76)	(−1.08)	(−1.72)	(−12.45)		
(22)	USE	Sell fee	−0.044	−0.044	−0.031	−0.689	−0.361	−1.184	−0.279	0.312	1,734
			(−0.87)	(−0.61)	(−2.64)	(−0.63)	(−0.98)	(−2.62)	(−8.81)		
(23)	USE	Legal fee	−0.290	0.296	−0.024	−0.777	1.593	0.418	−0.537	0.451	1,400
			(−4.20)	(3.63)	(−1.16)	(−1.04)	(2.92)	(0.78)	(−15.63)		
(24)	USE	Acct fee	−0.186	0.099	−0.094	−0.017	1.304	−0.307	−0.356	0.238	1,401
			(−2.19)	(1.14)	(−2.39)	(−0.02)	(2.67)	(−0.64)	(−11.97)		
(25)	MDA	Spread	−0.298	0.388	−0.052	2.631	−0.526	−0.747	−0.588	0.517	1,748
			(−3.01)	(2.56)	(−3.81)	(2.82)	(−1.66)	(−1.59)	(−11.91)		
(26)	MDA	Mgmt fee	−0.139	0.384	−0.013	−0.972	−0.258	−0.546	−0.177	0.381	1,650
			(−2.15)	(6.02)	(−1.80)	(−2.07)	(−1.69)	(−2.72)	(−9.48)		
(27)	MDA	Und. fee	−0.050	0.174	0.004	−0.894	−0.144	−0.416	−0.202	0.412	1,648
			(−0.63)	(2.97)	(0.64)	(−1.77)	(−1.02)	(−1.63)	(−12.26)		
(28)	MDA	Sell fee	0.082	0.046	−0.032	−0.713	−0.350	−1.181	−0.281	0.312	1,734
			(0.76)	(0.49)	(−2.65)	(−0.64)	(−0.97)	(−2.67)	(−9.21)		
(29)	MDA	Legal fee	−0.532	0.784	0.003	−0.810	1.631	0.416	−0.520	0.459	1,400
			(−4.14)	(3.71)	(0.16)	(−1.04)	(3.01)	(0.77)	(−15.66)		
(30)	MDA	Acct fee	−0.018	0.676	−0.084	0.051	1.398	−0.197	−0.347	0.243	1,401
			(−0.08)	(2.15)	(−2.22)	(0.05)	(2.82)	(−0.38)	(−12.30)		

OLS regressions with yearly fixed effects are presented for IPOs issued in the United States from January 1996 to October 2005, excluding firms with an issue price less than $5, ADRs, financial firms, unit IPOs, dual class IPOs, and REITs. The observations in each specification vary due to availability of fee data and are noted in the final column. The dependent variable is either the Underwriting spread (or its components including the Management fee, the Underwriting fee, and the Selling fee), the Issuer's legal fees, or the Issuer's accounting fees (all are expressed as a percentage of issue proceeds). We report results for the prospectus as a whole (TOTAL) and all four sections: Prospectus Summary (PS), Risk Factors (RF), Use of Proceeds (USE), and MD&A. Three key independent variables are the IPO's standard and informative content, which are based on the following first-stage regression for each IPO i: $norm_{tot,i} = a_{rec,i}\, norm_{rec,i} + a_{ind,i}\, norm_{ind,i} + \epsilon$. Standard content is the sum of the coefficients $a_{rec,i}$ and $a_{ind,i}$, and Informative content is the sum of the absolute residuals. The remaining independent variables are as follows. Log firm age is the log of the IPO year minus the firm's founding date, where founding dates are obtained from the Field–Ritter dataset, as used in Field and Karpoff (2002) and Loughran and Ritter (2004). UW $ market share is the lead underwriter's dollar market share in the past calendar year. UW law $ market share is the underwriting firm's legal counsel's dollar market share in the past calendar year. Issuer law $ market share is the issuer firm's legal counsel's dollar market share in the past calendar year. The Log issue size is the log of the dollar issue proceeds. Year and Industry fixed effects are also included, where industry definitions are based on the Fama–French forty-eight industry codes. t-Statistics are in parentheses and are adjusted for clustering within industries and years.

prospectus. Although due diligence can affect underwriting risk and the compensation needed for selling shares, its impact is likely to be less direct.

Table 8 presents results for each component of the spread as well as for the spread as a whole, and the results are striking. While the gross spread is increasing in both standard and informative content, the increase in each is due to different underwriting activities. For example, only selling fees are significantly related to standard content. This is consistent with our conjecture that standard content reflects less premarket information gathering and therefore requires a greater selling effort by the syndicate during bookbuilding.

In contrast, the greater the informative content, the higher are fees more directly related to due diligence, especially the management fee. The increase is also economically meaningful. A 1 SD increase in informative content results in a 4% increase in the average management fee and a 2% increase in the average underwriting fee. Informative content is generally unrelated to selling fees except in the Prospectus Summary.

Examining various sections of the prospectus sheds additional light on the role of information in expenses. Greater informative content in the Prospectus Summary affects only the selling fee. If the Prospectus Summary is used as a marketing tool, increasing the amount of informativeness reduces the effort needed to sell the document. The strongest relation between content and fees is in MD&A. The greater the informative (standard) content in MD&A, the higher (lower) are the expenses associated with the offer. Overall, these results on issuer expenses are consistent with differences in the level of effort to gather premarket information through due diligence and not simply due to differences in disclosure strategies.

5. Underwriter Content

Table 2 indicates that two IPOs brought to market by the same underwriter have content that is closer in similarity than two IPOs brought to market by different underwriters. This reflects an underwriter's "signature" in the document that may be unique to a specific underwriter, and we are interested in whether this signature contributes to or detracts from pricing accuracy. In this section, we examine the marginal contribution of unique underwriter content on IPO pricing.

We define unique underwriter content in the following manner: each IPO i has U IPOs filed by any of its lead underwriters preceding its initial filing, whose word vectors are denoted by $words_{tot,u}$. We then normalize this vector by dividing by the sum of its elements to get $norm_{tot,u}$. We exclude IPOs from the set of U IPOs if they were filed in the past ninety days or if they are in the same industry. This ensures that our three content types contain distinct information. Next, we define unique underwriter content ($norm_{uw,i}$) as

$$norm_{uw,i} = \frac{1}{U} \sum_{u=1}^{U} norm_{tot,u}.$$

Because unique underwriter content remains highly correlated with recent and industry IPO content even though we construct this variable using distinct IPOs, we orthogonalize underwriter content by taking the residual of its projection onto recent issued IPO ($norm_{rec,i}$) and industry content ($norm_{ind,i}$). The resulting residual is then normalized by dividing by the absolute value of its elements. We denote this as $norm(orth)_{uw,i}$, which represents content that is unique to the underwriter.

We run an extended version of the first-stage regression for each IPO:[30]

$$norm_{tot,i} = a_{rec,i}\, norm_{rec,i} + a_{ind,i}\, norm_{ind,i} + a_{uw,i}\, norm(orth)_{uw,i} + \epsilon. \quad (3)$$

In the underwriter extended model, standard content is still defined as the sum of the recent and industry coefficients as in Equation (2), but now, the regression also includes unique underwriter content ($a_{uw,i}$). Informative content is then the summed absolute residuals from Equation (3) rather than Equation (1).[31]

The marginal contribution of unique underwriter content on IPO pricing is presented in table 9. Although the coefficients on standard and informative content, even when including unique underwriter content, are relatively similar to those documented previously, greater unique underwriter content is associated with greater pricing accuracy. The higher the underwriter signature in the document, the greater are the reductions in ΔP and subsequent underpricing. This result is somewhat surprising given that underwriter content is estimated in a manner similar to standard content. Specific underwriter content, even though it exists in some past IPOs underwritten by the same lead underwriters, is clearly interpreted by investors as being informative. One possible explanation for why unique underwriter content affects pricing in this way is that greater underwriter content may reflect more involvement by the underwriter in drafting the initial prospectus and may be seen as certification of the issue by the underwriter.

The last two columns of the table split the sample into high- and low-reputation underwriters based on the median dollar market share. Classification into high- and low-underwriter reputation is conducted on a yearly basis. The increased accuracy in premarket pricing, as indicated by a lower change in offer price and initial return, is dominated by high-reputation underwriters. There are at least two possible explanations for this. First, high-reputation underwriters may have the necessary experience to assess the IPO relative to their low-reputation counterparts. In addition, high-reputation underwriters may be in a better position to credibly convey information. Second, it is possible that

[30] The sample size with underwriter content is slightly reduced to 1,666 IPOs, because we discard any IPO that does not have at least one past underwriter IPO that can be used in estimating Equation (3).

[31] In this specification, standard and informative content are −1.7% correlated, and orthogonalized underwriter content is +5.3% correlated with informative (residual) content and −20.9% correlated with standard text.

Table 9
Relation of underwriter's unique content to price adjustments and initial returns

Row	Section	Dependent variable	Standard content	Informative content	Unique UW content	Year+Ind eff. + controls	Unique high $ UW content	Unique low $ UW content
(1)	TOTAL	ΔP	0.214	−0.282	−0.184	Yes	−0.204	−0.097
			(1.85)	(−2.20)	(−3.49)		(−2.90)	(−1.01)
(2)	TOTAL	abs(ΔP)	0.022	−0.398	−0.108	Yes	−0.183	−0.024
			(0.26)	(−3.97)	(−3.06)		(−3.12)	(−0.37)
(3)	TOTAL	IR	0.531	−1.015	−0.498	Yes	−0.656	−0.145
			(1.99)	(−3.28)	(−3.35)		(−3.03)	(−1.03)
(4)	PS	ΔP	0.165	0.053	−0.101	Yes	−0.119	−0.055
			(2.65)	(0.58)	(−3.23)		(−2.43)	(−1.33)
(5)	PS	abs(ΔP)	0.088	0.056	−0.030	Yes	−0.045	−0.025
			(2.17)	(0.69)	(−0.92)		(−0.97)	(−0.70)
(6)	PS	IR	0.402	0.136	−0.251	Yes	−0.386	−0.062
			(3.82)	(0.91)	(−2.93)		(−3.57)	(−0.83)
(7)	RF	ΔP	−0.008	−0.149	−0.160	Yes	−0.168	−0.100
			(−0.13)	(−2.26)	(−4.50)		(−3.50)	(−1.68)
(8)	RF	abs(ΔP)	0.110	−0.262	−0.069	Yes	−0.127	−0.013
			(2.00)	(−5.31)	(−2.86)		(−3.10)	(−0.28)
(9)	RF	IR	0.197	−0.509	−0.311	Yes	−0.417	−0.092
			(0.93)	(−3.00)	(−3.67)		(−3.13)	(−0.87)
(10)	USE	ΔP	0.092	−0.016	−0.021	Yes	−0.033	0.029
			(3.17)	(−0.46)	(−1.31)		(−1.45)	(0.86)
(11)	USE	abs(ΔP)	0.099	−0.069	−0.009	Yes	−0.022	0.015
			(4.72)	(−2.82)	(−0.73)		(−1.19)	(0.76)
(12)	USE	IR	0.329	−0.181	−0.109	Yes	−0.146	−0.007
			(5.19)	(−1.94)	(−3.31)		(−2.89)	(−0.11)
(13)	MDA	ΔP	0.112	−0.298	−0.107	Yes	−0.068	−0.104
			(2.57)	(−4.23)	(−2.94)		(−1.58)	(−2.32)
(14)	MDA	abs(ΔP)	0.055	−0.187	−0.028	Yes	−0.041	−0.018
			(1.58)	(−2.95)	(−1.04)		(−0.85)	(−0.57)
(15)	MDA	IR	0.358	−0.838	−0.280	Yes	−0.336	−0.172
			(3.10)	(−5.27)	(−4.79)		(−3.66)	(−2.55)

OLS regressions with yearly fixed effects are presented for 1,666 IPOs issued in the United States from January 1996 to October 2005, excluding firms with an issue price less than $5, ADRs, financial firms, unit IPOs, dual class IPOs, and REITs. All but the last two columns display results for the model based on this entire sample. The last two columns report results only for the unique underwriter content variable when the model is run on the above median and below median underwriter market share subsamples. The dependent variable in all cases is either the Change in price from the filing date midpoint to the IPO offer price (ΔP), the Absolute value of the change in offer price (abs(ΔP)), or the Initial return (IR). We report results for the prospectus as a whole (TOTAL) and all four sections: Prospectus Summary (PS), Risk Factors (RF), Use of Proceeds (USE), and MD&A. Three key independent variables are the IPO's standard, informative, and unique underwriter content, which are based on the following first-stage regression for each IPO i: $norm_{tot,i} = a_{rec,i}\ norm_{rec,i} + a_{ind,i}\ norm_{ind,i} + a_{uw,i}\ norm(orth)_{uw,i} + \epsilon$. Standard content is the sum of the coefficients $a_{rec,i}$ and $a_{ind,i}$, Unique UW content is the coefficient $a_{uw,i}$, and Informative content is the sum of the absolute residuals. The remaining independent variables are as follows but are not shown to conserve space. Log firm age is the log of the IPO year minus the firm's founding date, where founding dates are obtained from the Field–Ritter dataset, as used in Field and Karpoff (2002) and Loughran and Ritter (2004). UW $ market share is the lead underwriter's dollar market share in the past calendar year. UW law $ market share is the underwriting firm's legal counsel's dollar market share in the past calendar year. Issuer law $ market share is the issuer firm's legal counsel's dollar market share in the past calendar year. The VC dummy is equal to 1 if a firm is VC financed. Auditor $ market share is the auditor's dollar market share in the past calendar year. Pre-file Nasdaq return is the NASDAQ return for the thirty trading days preceding the filing date. The Log dollars filed is the log of the original filing amount. Tech dummy is based on Loughran and Ritter (2004). Year and Industry fixed effects are also included, where industry definitions are based on the Fama–French forty-eight industry codes. The t-Statistics are in parentheses and are adjusted for clustering within industries and years.

these results reflect the fact that a precise estimation of unique underwriter content is more difficult for low-reputation underwriters, because they bring fewer IPOs to market.

Note that our findings are unlikely to be due to the type of IPOs brought to market by high-reputation underwriters, as table 5 documents a positive relation between the reputation of the underwriter and IPO pricing. Thus, our results may shed light on the puzzling phenomenon of the time-dependent change in the relation between underwriter reputation and underpricing, which has called into question the "certification" hypothesis of underwriter services. Although more recent issues brought to market by high-reputation underwriters have greater underpricing than their low-reputation counterparts, our results highlight that high-reputation underwriters can influence the amount of underpricing through enhanced due diligence and disclosure.

6. Topical and Tone Content

The prior sections noted a relation between standard and informative content and IPO pricing. This section examines whether the topic or tone of disclosure matters in terms of pricing. Further, topical content sheds additional light on the different roles that different parts of the prospectus play in information generation.

In order to examine the influence of topical content, we first compile word vectors or lists that relate to specific topics whose content and source are detailed in Appendix B.[32] For example, consider our legal jargon word list as the list of all of the words in the legal glossary "www.learnaboutlaw.com." Our starting point is a word vector from this Web site ($words_{legal}$) and its normalized version $norm_{legal}$. Note that these word vectors do not have an i subscript for IPO i, as content and tone word lists are universal. We construct similar word vectors for each word list.

[32] We decided to use existing word lists, to be conservative in our method, although it is possible to create word lists from existing text. We did examine the following word lists, which either had no effect or were deemed to be of secondary importance and therefore are not shown to conserve space: bond (http://www.investopedia.com/categories/bonds.aspcomptxt), competition (http://www.concurrences.com/rubrique.php3?id_rubrique=161), political economy (http://www.auburn.edu/johnspm/gloss/competition), international (http://www.importexporthelp.com/a/b2b-definitions.htm#A), government contract (http://www.targetgov.com/Content.asp?id=2409), Labor Union (//org.teamster.org/glossary.htm), merger and acquisition (http://www.investorsedge.com/dictionary/Mergers_and_Acquisitions_dictionary_category.html), ethics (http://www.ethics.org/resources/ethics-glossary.asp), and underwriting (http://www.investopedia.com/categories/ipos.asp). We use only one word list to capture a specific topic and did not explore whether any other similar word lists would improve the statistical significance. While we try to use lists that appear to be comprehensive in their choice of words, we do not supplement or change the list in any way except the Product Market word list, where we exclude financial, accounting, and legal terms. A word list, therefore, may have an insignificant presence or impact, because either it does not appropriately capture the specific content or it truly has no effect. Thus, we are careful not to interpret an insignificant result as indication that certain topics are unimportant.

We decompose the total word vector of each IPO prospectus into its exposure to each of the eight word lists and two tone lists as follows:

$$norm_{tot,i} = \sum_{c=1}^{C} a_{con,c,i}\, norm_{con,c} + \sum_{t=1}^{T} a_{tone,t,i}\, norm_{tone,t} + \epsilon. \quad (4)$$

Here, C is the number of specific content word lists considered (we use eight specific content lists), and T is the number of tone word lists considered (we use two specific tone lists). In this specification, we no longer utilize information from past IPOs and are, therefore, able to increase the sample size from 1,750 to 1,913 IPOs.

Table 10 presents summary statistics on the average coefficients on topical content and tone. The average IPO prospectus has higher relative exposure to accounting, marketing, and corporate governance content. Not surprisingly, the tone of the prospectus is also more positive than negative.

More interesting, however, is how topical content relates to the different role each section plays in the prospectus. For example, the Prospectus Summary has high exposure to marketing words, which is consistent with its hypothesized

Table 10
Summary statistics on topical content

Content type	Total document	Prospectus Summary	Risk Factors	Use of Proceeds	MD&A
Product market	0.032	0.049	0.024	−0.017	0.032
	(2.357)	(2.199)	(1.373)	(−0.374)	(1.429)
Accounting	0.124	0.159	0.082	0.203	0.239
	(12.247)	(9.483)	(6.310)	(6.435)	(14.308)
Legal	0.040	0.010	0.031	0.022	0.022
	(5.655)	(0.920)	(3.479)	(1.076)	(2.006)
Corporate strategy	0.015	0.022	0.008	0.133	0.056
	(3.591)	(3.260)	(1.561)	(10.136)	(8.117)
Patent	0.057	0.060	0.055	−0.011	0.034
	(6.503)	(4.058)	(4.893)	(−0.312)	(2.388)
Marketing	0.072	0.094	0.074	0.086	0.077
	(8.191)	(6.366)	(6.513)	(3.080)	(5.290)
Corp. governance	0.068	0.040	0.054	−0.008	−0.005
	(7.446)	(2.618)	(4.604)	(−0.282)	(−0.244)
Corporate valuation	0.040	0.040	0.063	0.085	0.082
	(5.647)	(3.544)	(7.019)	(3.917)	(7.121)
Positive tone	0.073	0.069	0.089	0.039	0.064
	(4.012)	(2.316)	(3.868)	(0.732)	(2.207)
Negative tone	0.021	0.008	0.059	−0.007	0.026
	(1.309)	(0.313)	(2.929)	(−0.132)	(1.005)

Summary statistics on the coefficients from the first-stage regressions on document content for 1,913 IPOs issued in the United States from August 1996 to October 2005, excluding firms with an issue price less than $5, ADRs, financial firms, unit IPOs, dual class IPOs, and REITs. The mean coefficients on the topical ($a_{con,c,i}$) and tone ($a_{tone,t,i}$) word lists are from the first-stage regression for each IPO i: $norm_{tot,i} = \sum_{c=1}^{C} a_{con,c,i}\, norm_{con,c} + \sum_{t=1}^{T} a_{tone,t,i}\, norm_{tone,t} + \epsilon$, , where C is the number of topical word lists, and T is the number of tone word lists. One regression is run for each document as a whole and for each section. For additional information on word and tone lists, see Appendix B. Average t-statistics over the full sample of IPOs from the first-stage regressions are in parentheses.

role as a marketing tool used by underwriters to attract investors. It also has high exposure to accounting content and is positive in tone.

The Risk Factors section has the highest exposure to accounting, marketing, and valuation and is the only section that has significant loadings on both positive and negative tone. This is consistent with its role in describing the second moment of outcomes (high and low), as its namesake suggests it should.

The Uses of Proceeds section has high exposure to accounting and strategy, as would be expected, since this section is an explanation of how the proceeds from the IPO are to be used by the issuing firm.

Finally, MD&A has strong exposure to accounting, valuation, and strategy, which conforms to its role as management's assessment of both the past performance of the firm and its potential future prospects. This section also has significant positive tone and a high exposure to marketing words, which may reflect managerial optimism.

Table 11 explores, in further detail, whether specific topics or tone are related to subsequent pricing. Note that the same control variables as in table 5 are included in the regression but are not reported to conserve space.

The overall results provide a strong indication that topical content related to valuation and due diligence is associated with a reduction in both price changes during bookbuilding and the initial return. Thus, increased disclosure on both tangible (accounting and product market) and intangible information (corporate strategy) appears to be correlated with greater accuracy in pricing. Content associated with marketing, however, has the opposite effect. While Chemmanur and Yan (2008) find that firms that engage in more product marketing prior to the IPO have lower underpricing, our results suggest that firms that talk about marketing are more likely to have information revealed during bookbuilding and have higher initial returns. This may be due to the use of marketing words to hype the issue.

Using the Harvard Inquirer Categories (Tetlock 2007), we find a marginal effect on tone for the document as a whole, which appears to be driven by the Risk Factors section. This is consistent with the role of this section in assessing and mitigating liability risk. Since the underwriter and issuer are liable, both legally and reputationally, for any misstatements in the prospectus, a net positive tone sends a strong signal to investors regarding the expected riskiness and valuation of the issue. This is associated with a reduction in the change in the offer price and the initial return.

7. Conclusion

We examine whether a trade-off exists between pricing an issue using information gathered from premarket due diligence and information gathered from investors during bookbuilding by decomposing the text of the initial prospectus into standard (content in recent and past industry IPOs) and informative (residual content) components. We are interested in whether there exists an

Table 11
Relation of word content and tone to price adjustments and initial returns

Row	Section	Dependent variable	Prod. mkt	Accting	Legal	Corp. strat.	Patent & trade	Marketing	Corp govern.	Valuation	Net tone	Year+Ind eff. + controls	R^2
(1)	TOTAL	ΔP	−0.554	−0.567	−0.187	−1.253	−0.651	1.147	0.223	0.734	−0.257	Yes	0.205
			(−1.76)	(−1.78)	(−0.35)	(−1.95)	(−1.29)	(2.70)	(0.50)	(1.29)	(−0.89)		
(2)	TOTAL	abs(ΔP)	0.186	−0.713	−0.104	−1.062	−0.151	1.024	0.111	−0.178	−0.178	Yes	0.167
			(0.96)	(−4.02)	(−0.33)	(−1.73)	(−0.32)	(3.30)	(0.48)	(−0.40)	(−0.86)		
(3)	TOTAL	IR	−1.174	−1.207	−0.834	−4.976	−0.326	2.918	−0.512	0.870	−1.418	Yes	0.256
			(−1.70)	(−1.30)	(−0.82)	(−3.17)	(−0.24)	(3.47)	(−0.44)	(0.74)	(−1.86)		
(4)	PS	ΔP	−0.143	0.125	0.519	−0.431	−0.230	0.484	0.362	−0.013	−0.046	Yes	0.202
			(−0.67)	(1.15)	(1.95)	(−1.46)	(−0.62)	(2.71)	(1.74)	(−0.04)	(−0.30)		
(5)	PS	abs(ΔP)	0.103	−0.098	0.096	−0.282	−0.165	0.436	0.137	−0.136	−0.174	Yes	0.158
			(0.84)	(−1.21)	(0.36)	(−1.39)	(−0.62)	(2.79)	(0.99)	(−0.59)	(−1.54)		
(6)	PS	IR	−0.380	−0.095	0.869	−0.996	−0.569	0.994	−0.262	−0.860	−0.430	Yes	0.249
			(−0.98)	(−0.35)	(0.96)	(−1.59)	(−0.98)	(2.84)	(−0.65)	(−1.25)	(−1.15)		
(7)	RF	ΔP	−0.341	−0.781	0.077	−0.867	−0.242	0.269	−0.414	−0.338	−0.665	Yes	0.204
			(−1.00)	(−2.81)	(0.21)	(−1.58)	(−0.66)	(0.93)	(−1.48)	(−0.96)	(−3.56)		
(8)	RF	abs(ΔP)	0.086	−0.524	0.027	−0.227	0.427	0.564	−0.763	−0.345	−0.280	Yes	0.166
			(0.39)	(−2.79)	(0.11)	(−0.68)	(1.54)	(2.58)	(−3.93)	(−1.70)	(−1.90)		
(9)	RF	IR	−0.784	−0.864	0.560	−3.995	0.728	1.821	−2.542	0.141	−1.528	Yes	0.260
			(−1.60)	(−1.57)	(0.74)	(−4.01)	(0.65)	(2.57)	(−2.37)	(0.15)	(−4.23)		

(continued)

Table 11
Continued

(10)	USE	ΔP	−0.252	−0.198	−0.246	0.094	−0.163	−0.114	−0.114	0.327	0.113	Yes	0.204
			(−1.10)	(−2.72)	(−1.59)	(0.83)	(−1.27)	(−1.09)	(−0.96)	(2.51)	(1.00)		
(11)	USE	abs(ΔP)	0.157	−0.068	−0.094	0.215	−0.073	0.143	−0.049	0.302	0.022	Yes	0.162
			(1.20)	(−1.34)	(−1.04)	(2.63)	(−0.64)	(1.59)	(−0.57)	(3.09)	(0.31)		
(12)	USE	IR	−0.544	−0.380	−0.633	0.322	−0.155	0.224	0.127	1.592	0.231	Yes	0.263
			(−1.30)	(−2.91)	(−2.23)	(1.41)	(−0.39)	(1.07)	(0.63)	(2.60)	(0.92)		
(13)	MDA	ΔP	−0.225	−0.608	−0.459	−0.225	0.004	0.098	−0.204	0.099	−0.182	Yes	0.205
			(−0.82)	(−3.55)	(−1.62)	(−1.09)	(0.01)	(0.48)	(−0.64)	(0.33)	(−0.93)		
(14)	MDA	abs(ΔP)	−0.188	−0.242	−0.101	0.007	−0.042	0.292	−0.328	−0.568	−0.126	Yes	0.161
			(−1.39)	(−3.22)	(−0.63)	(0.03)	(−0.19)	(2.57)	(−1.76)	(−4.68)	(−0.87)		
(15)	MDA	IR	−0.752	−1.175	−0.846	−1.259	0.160	0.860	−1.804	−0.337	−0.716	Yes	0.255
			(−1.08)	(−2.75)	(−1.82)	(−2.30)	(0.26)	(2.38)	(−1.95)	(−0.52)	(−1.16)		

OLS regressions with yearly fixed effects are presented for 1,913 IPOs issued in the United States from January 1996 to October 2005, excluding firms with an issue price less than $5, ADRs, financial firms, unit IPOs, dual class IPOs, and REITs. The dependent variable is either the Change in price from the filing date midpoint to the IPO offer price (ΔP), the Absolute value of the change in offer price (abs(ΔP)), or the Initial return (IR). We report results for the prospectus as a whole (TOTAL) and all four sections: Prospectus Summary (PS), Risk Factors (RF), Use of Proceeds (USE), and MD&A. The key independent variables are the coefficients on the topical ($a_{con,c,i}$) and tone ($a_{tone,t,i}$) word lists (described in Appendix B) from the first-stage regression for each IPO i: $norm_{tot,i} = \sum_{c=1}^{C} a_{con,c,i}\ norm_{con,c} + \sum_{t=1}^{T} a_{tone,t,i}\ norm_{tone,t} + \epsilon$, where C is the number of topical word lists (C=8), and T is the number of tone word lists (T=2). Net tone is the difference in the positive and negative tone word exposures. Other remaining independent variables include year and industry fixed effects, in addition to the same control variables that are presented in table 5; however, these variables are not displayed to conserve space. t-Statistics are in parentheses and are adjusted for clustering within industries and years.

alternate mechanism for issuers and underwriters to mitigate the potentially high initial returns associated with bookbuilding. We hypothesize that greater informative content should improve pricing accuracy, as measured by a lower absolute change in the offer price and lower initial returns. In contrast, greater standard content implies more reliance on investors to price the issue during bookbuilding, resulting in higher changes in offer prices and higher initial returns.

Our results support these hypotheses. Greater informative (standard) content decreases (increases) both the price change from the filing midpoint to the IPO price and underpricing. The reduction is economically meaningful, indicating that premarket information production can significantly increase pricing accuracy and reduce required information rents paid during bookbuilding. The largest improvement in pricing accuracy is associated with informative content in the management's discussion section. Our results suggest that the management's role in premarket due diligence and information generation might be significant, a point overlooked by many studies of IPO pricing.

We also find evidence consistent with differences in the amount of effort expended in the premarket. Lawyer and auditor fees are significantly higher when there is more informative content in the initial prospectus. A decomposition of the gross spread indicates that the management fee, the component most likely to reflect underwriter effort in due diligence, is also increasing in the amount of informative content. The only component of the gross spread affected by standard content is the selling fee. This supports the conjecture that IPOs with more standard content require greater selling effort during bookbuilding. These relationships support the existence of a trade-off between greater effort in premarket due diligence and costly bookbuilding.

Note, however, that our results cannot determine whether issuing firms that choose to avoid premarket information gathering are acting suboptimally. There are costs and benefits to collecting information prior to filing that are likely to vary across firms. If all firms are acting optimally, then firms with higher information costs will collect less.[33] Additional research is needed to determine which firms would benefit most from greater information gathering in the premarket.

We also examine the role of unique underwriter content on pricing accuracy and find that unique underwriter content also improves pricing accuracy. This reduction, however, is significant for only underwriters with greater market share and, thus, higher reputation and experience. Finally, we explore the type of informative content that is most significantly related to pricing and find that content directly related to inputs into valuation models most likely used by practitioners seems to matter most.

Given that Jenkinson and Jones (2009) find that only about half of their sample respondents build valuation models for evaluating IPOs, our results suggest

[33] We thank the referee for pointing this out.

that the passive evaluation of IPO prospectuses using text analysis may be useful to investors. Although we differ in the extent to which we believe investors provide information during bookbuilding, we concur with their interpretation that information "flow is likely to include information from underwriter to investor as well as vice versa."

Appendix A

This Appendix explains how we compute the document similarity between two documents i and j. This same procedure can be applied to document sections, in which case the result would be the "section similarity."

We first take the text in each document (or document section) and construct a numerical vector summarizing the counts of its English-language word roots. This vector has a number of elements equal to the number of word roots, and one element is the number of times the given word root appears in the document. Word roots are identified by Webster.com, and we use a Web crawling algorithm to build a database of the unique word roots that correspond to all English-language words that appear in the universe of all IPO prospectuses. For example, the words display, displayed, and display all have the same word root "display."[34] In order to conserve computing space, we exclude articles, conjunctions, personal pronouns, abbreviations, compound words, and any words that appear fewer than a total of five times in the universe of all words from these counts, because they are not informative regarding content. This leaves a vector of 5,803 possible words. We define this vector for the document as $words_{tot,i}$ ($words_{ps,i}$, $words_{rf,i}$, $words_{use,i}$, and $words_{mda,i}$ for sections), the total number of root words used.

To measure the degree of similarity of documents i and j, we simply take the dot product of the two word vectors normalized by their vector lengths. This quantity is widely used in studies of information processing and is known as the "cosine similarity" method (see Kwon and Lee 2003 for more information), because it measures the angle between two word vectors on a unit sphere. We refer to this quantity as document similarity, and we utilize this measure in Section 3.

$$Document\ Similarity_{tot,i,j} = \frac{words_{tot,i} \cdot words_{tot,j}}{\|words_{tot,i}\| \ \|words_{tot,j}\|}. \tag{A1}$$

Because all word vectors $words_{tot,i}$ have elements that are nonnegative, this measure of document similarity has the nice property of being bounded in the interval [0,1]. Intuitively, the similarity between two documents (or word lists) is closer to 1 when they are more similar and can never be less than 0 if they are entirely different.

Appendix B

We use the following word lists to assess the type of topical content:

Product Market: All words appearing in the Standard Industrial Classification code industry definitions as provided by the SEC (excluding financial, accounting, and legal terms).

Accounting: All words appearing in the COMPUSTAT data item list.

Legal: Words from the following legal glossary:
http://www.learnaboutlaw.com/General/glossary.htm

[34] Methodologically, we first create a vector of all word counts in the given section of the document, and we then replace each word with its word root. We then tabulate the frequency vector for the given document section based on the total counts of each word root.

Corporate Strategy: Merged universe of words from the glossary of the Hill and Jones textbook and the Ross, Westerfield, and Jaffe textbook table of contents.

Patent: Merged universe of words from the patent glossary: http://www.bpmlegal.com/patgloss.html, and the intellectual property/trademarks glossary: http://marklaw.com/trademark-glossary/glossary.htm

Marketing: Words from the following marketing glossary: http://marketing.about.com/od/marketingglossary/a/marketingterms.htm

Valuation: Words from the following valuation methods Web site: http://fvs.aicpa.org/Resources/Business+Valuation/Tools+and+Aids/Definitions +and+Terms/International+Glossary+of+Business+Valuation+Terms.htm

Corporate Governance: Words from the following corporate governance glossary: www.corp-gov.org/glossary.php3

We also consider the following tone word lists, which are analogously defined.

Negative: List of negative words from: http://www.wjh.harvard.edu/ inquirer/homecat.htm

Positive: List of positive words from: http://www.wjh.harvard.edu/ inquirer/homecat.htm

References

Arnold, T., R. P. Fishe, and D. North. 2008. The Effects of Ambiguous Information on Initial and Subsequent IPO Returns. Working Paper, University of Richmond.

Barry, C., C. Muscarella, J. Peavy, and M. Vetsuypens. 1990. The Role of Venture Capital in the Creation of Public Companies. *Journal of Financial Economics* 27:447–71.

Beatty, R., and J. Ritter. 1986. Investment Banking, Reputation and the Underpricing of Initial Public Offerings. *Journal of Financial Economics* 15:213–32.

Beatty, R., and I. Welch. 1996. Issuer Expenses and Legal Liability in Initial Public Offerings. *Journal of Law and Economics* 39:545–602.

Benveniste, L., and P. Spindt. 1989. How Investment Bankers Determine the Offer Price and Allocation of New Issues. *Journal of Financial Economics* 24:343–62.

Bhattacharya, S., and G. Chiesa. 1995. Proprietary Information, Financial Intermediation, and Research Incentives. *Journal of Financial Intermediation* 4:328–57.

Bhattacharya, S., and J. Ritter. 1983. Innovation and Communication: Signaling with Partial Disclosure. *Review of Economic Studies* 50:331–46.

Boukus, E., and J. Rosenberg. 2006. The Information Content of FOMC Minutes. Working Paper, Yale University.

Chemmanur, T., and A. Yan. 2008. Product Market Advertising and New Equity Issues. *Journal of Financial Economics* 92:40–65.

Chen, H. C., and J. Ritter. 2000. The Seven Percent Solution. *Journal of Finance* 55:1105–31.

Cook, D., R. Kieschnick, and R. Van Ness. 2006. On the Marketing of IPOs. *Journal of Financial Economics* 82:35–61.

Cornelli, F., and D. Goldreich. 2003. Bookbuilding: How Informative Is the Order Book? *Journal of Finance* 58:1415–43.

Darrough, M. N., and N. M. Stoughton. 1990. Financial Disclosure Policy in an Entry Game. *Journal of Accounting and Economics* 12:219–43.

Dye, R. A. 2001. An Evaluation of "Essays on Disclosure" and the Disclosure Literature in Accounting. *Journal of Accounting and Economics* 32:181–235.

Ertimur, Y., and M. Nondorf. 2009. SEC Comment Letters for IPO Firms. Working Paper, Duke University.

Field, L. C., and J. Karpoff. 2002. Takeover Defenses of IPO Firms. *Journal of Finance* 57:1857–89.

Guo, R. J., B. Lev, and N. Zhou. 2004. Competitive Costs of Disclosure by Biotech IPOs. *Journal of Accounting Research* 42:319–64.

Hanley, K. W. 1993. The Underpricing of Initial Public Offerings and the Partial Adjustment Phenomenon. *Journal of Financial Economics* 34:231–50.

Healy, P. M., and K. G. Palepu. 2001. Information Asymmetry, Corporate Disclosure, and the Capital Markets: A Review of the Empirical Disclosure Literature. *Journal of Accounting and Economics* 31:405–40.

Hoberg, G., and G. Phillips. 2008. Product Market Synergies and Competition in Mergers and Acquisitions. Working Paper, University of Maryland.

Jenkinson, T., and H. Jones. 2004. Bids and Allocations in European IPO Bookbuilding. *Journal of Finance* 59:2309–38.

———. 2009. IPO Pricing and Allocation: A Survey of the Views of Institutional Investors. *Review of Financial Studies* 22:1477–1504.

Khanna, N., T. Noe, and R. Sonti. 2008. Good IPOs Draw in Bad: Inelastic Banking Capacity in the Primary Issue Market. *Review of Financial Studies* 21:1873–1906.

Kim, M., and J. Ritter. 1999. Valuing IPOs. *Journal of Financial Economics* 53:409–37.

Kwon, O. W., and J. H. Lee. 2003. Text Categorization Based on *k*-Nearest Neighbor Approach for Web Site Classification. *Information Processing & Management* 39:25–44.

Leone, A. J., S. Rock, and M. Willenborg. 2007. Disclosure of Intended Use of Proceeds and Underpricing of Initial Public Offerings. *Journal of Accounting Research* 45:111–53.

Li, F. 2006. Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports? Working Paper, University of Michigan.

Liu, L., A. Sherman, and Y. Zhang. 2007. Media Coverage and IPO Pricing. Working Paper, Hong Kong University and University of Notre Dame.

Ljungqvist, A., and W. Wilhelm. 2003. IPO Pricing in the Dot-com Bubble. *Journal of Finance* 58:723–52.

Logue, D. 1973. On the Pricing of Unseasoned Equity Issues 1965–69. *Journal of Financial and Quantitative Analysis* 8:91–103.

Loughran, T., and B. McDonald. 2008. Plain English. Working Paper, Notre Dame University.

Loughran, T., and J. Ritter. 2002. Why Don't Issuers Get Upset About Leaving Money on the Table in IPOs. *Review of Financial Studies* 15:413–33.

———. 2004. Why Has IPO Underpricing Changed Over Time? *Financial Management* 33:5–37.

Lowry, M., and W. Schwert. 2002. IPO Market Cycles: Bubbles or Sequential Learning? *Journal of Finance* 57:1171–1200.

———. 2004. Is the IPO Pricing Process Efficient? *Journal of Financial Economics* 71:3–26.

Maksimovic, V., and P. Pichler. 2001. Technological Innovation and Initial Public Offerings. *Review of Financial Studies* 14:459–94.

Markov, A. A. 1913/2006. Classical Text in Translation: An Example of Statistical Investigation of the Text *Eugene Onegin* Concerning the Connection of Samples in Chains. *Science in Context* 19:591–600.

Megginson, W., and K. Weiss. 1991. Venture Capitalist Certification in Initial Public Offerings. *Journal of Finance* 46:879–903.

Mohan, S. 2007. Disclosure Quality and Its Effect on Litigation Risk. Working Paper, University of Texas.

Nelson, K., and A. C. Pritchard. 2008. Litigation Risk and Voluntary Disclosure: The Use of Meaningful Cautionary Language. Working Paper, Rice University.

Schrand, C., and R. Verrecchia. 2005. Information Disclosure and Adverse Selection Explanations for IPO Underpricing. Working Paper, Wharton School.

Sherman, A., and S. Titman. 2002. Building the IPO Order Book: Underpricing and Participation Limits with Costly Information. *Journal of Financial Economics* 65:3–29.

Spatt, C., and S. Srivastava. 1991. Preplay Commmunication, Participation Restrictions and Efficiency in Initial Public Offerings. *Review of Financial Studies* 4:709–26.

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

Tetlock, P., M. Saar-Tsechanksy, and S. Macskassy. 2008. More Than Words: Quantifying Language to Measure Firms' Fundamentals. *Journal of Finance* 63:1437–67.

Torstila, S. 2001. The Distribution of Fees within the IPO Syndicate. *Financial Management* 30:25–43.

Verrecchia, R. E. 2001. Essays on Disclosure. *Journal of Accounting and Economics* 32:97–180.