Internet Appendix to:

A Text-Based Analysis of Corporate Innovation

by Gustaf Bellstam, Sanjai Bhagat and J. Anthony Cookson

This Internet Appendix contains a comprehensive set of additional statistics and robustness exercises for our paper, "A Text-Based Analysis of Corporate Innovation."

Table A.1 provides a list of variable definitions.

Section A.2 presents several additional details on the LDA fit and robustness. Table A.2 presents a table of the t-statistics of a regression of patent counts on each fitted topic in the fitted LDA model. Figure A.1 presents word clouds of two non-innovation topics, as a basis of comparison to the main word cloud in the paper. Table A.3 presents a list of common words from the innovation topic, ordered by their relative frequency. Table A.4 presents evidence on the stability of innovation language over time, in relation to the benchmark textbook. Figure A.2 presents a comparison of the innovation topic versus other topics to an alternative textbook on innovation – Drucker (1985). Finally, we present some details on a 50-topic LDA fit: word clouds from two innovation topics (Figure A.3), and main results using the best topic from the 50-topic LDA (Table A.5).

Section A.3 of this appendix presents tables that include the full results (with controls reported) for the main specifications, as well as some notable alternative specifications. Table A.6 presents the full results. For the subset of patenting firms, Table A.7 implements a version of the main specification, but with a continuous measure of patenting intensity (logged patents) as the interactive variable. Table A.8 presents the main specification, but *without* controls for patent counts or citations, which provides context for the main tests. Table A.9 presents the specification with the interaction for non-patenting firm that also controls for patent counts and patent citations, as robustness to the main table, which does not include these patenting controls. Table A.10 presents the results on longer-term dynamics of the relation between text-based innovation and different measures of performance, which are summarized in the main paper in Figure 8. Finally, Table A.11 presents the main specification, but with standard errors double clustered by firm and year (as robustness to the main specifications, which include firm-level clustering).

Section A.4 presents several notable robustness tests that could not be included in the main paper for the sake of brevity. Table A.12 presents the main result, but restricting to firms with below-median analyst sentiment as an alternative control for sentiment. Table A.13 presents the main results, but restricting to firms in which analysts use above-median forward-looking language (based on a measure of forward-looking intensity adapted from Muslu et al., 2015). Table A.14 presents the main results from a specification that instruments for firm text-based innovation using industry-rival values of text-based innovation (to alleviate concerns about strategic disclosure). Table A.15 presents the main results, plus an interaction with the timing of the benchmark textbook. Table

A.16 presents the results from an analysis in which the original textual corpus is purged of words that begin with "gro" or "rev."

Section A.5 presents additional results on other innovation outcomes that were not included in the main text. Table A.17 presents results on how text-based innovation relates to patent value and product announcements one year ahead (date $t+1$). Table A.18 presents results on how "negative" text-based innovation relates to other innovation outcomes. Table A.19 presents results on how text-based innovation relates to R&D intensity, both contemporaneously and one year ahead.

Finally, Section B presents two extended discussions of alternative approaches for building the text-based innovation measure. Section B.1 describes our fourth-root scaling of the topic loadings (maintained throughout the main text) in comparison to using an inverse hyperbolic sine (approximate log) transformation. Section B.2 presents an "innovation" word list approach to building a text-based measure of corporate innovation, and contrasts the approach and findings with our topic modeling approach.

A Appendix Tables and Figures

A.1 Variable Definitions

Table A.1: Variable Definitions

Note: This table includes variable definitions and descriptions for outcome and control variables used throughout the paper. The data source is Compustat unless otherwise noted. As the main text includes a full discussion of the text-based innovation measure, the reader should refer to those sections for a description.

Variable	**Name**	**Description**
ROA	Return on assets	*EBITDA scaled by Total Assets*
Q	Tobin's Q	*Market value of equity plus total assets minus common equity and deferred taxes divided by total assets*
$Salesgrowth_t$	Sales growth	*The percentage change in sales in between year t and $t-1$ (decimal form)*
Tangibility	Asset tangibility	*Property plant and equipment divided by total assets*
Leverage	Leverage	*Total liabilities divided by assets, replacing book equity with market equity as of the last day of the fiscal year*
Age	Age	*The number of years since the first entered Compustat (earliest date 1975)*
Cash/Assets	Cash to assets ratio	*The ratio of cash to assets taken from Compustat for year t*
Patents	Patent count	*The number of patent applications in year t that correspond to an eventually granted patent*
Citations	Citation count	*The number of citations to patents applied for in year t*
Patenting Firm	Patenting Firm	*An indicator (=1) for whether a firm ever has a non-zero value of Patents.*
Patent Value	Patent Value	*The abnormal stock increase (in $millions) on the day of the granted patent (from Kogan et al. (2017))*
Products	Product Announcements	*The count of product amouncements in which the stock return exceeded the 75th percentile in Mukherjee, Singh, and Zaldokas (2016).*

A.2 Additional Detail on LDA

Table A.2: Fit of Patenting Outcomes to Loadings for Every Topic in the 15-Topic LDA

Note: This table presents the t-statistics and adjusted R-squared on the linear relationship between a firm's patent applications and the loadings of each of the 15 topics from the fitted LDA model. Topic 6 is the innovation topic that we use for our text-based measure of innovation. This topic explains nearly two times the variation in patenting that any other topic can explain, and the word distribution is closest to the word frequencies in an innovation textbook (Tidd, Bessant, and Pavitt, 2005). Errors are double clustered on firm and year.

Topic	T-Stat	Adj R^2
6	12.372	0.047
15	8.697	0.024
12	6.915	0.015
2	6.773	0.014
11	4.718	0.007
7	2.908	0.002
10	-0.534	-0.0002
1	-3.764	0.004
5	-5.246	0.009
8	-5.722	0.010
4	-7.361	0.017
3	-7.394	0.017
13	-7.646	0.018
9	-7.888	0.020
14	-8.678	0.024

Figure A.1: Word Clouds of Two Other Fitted Topics

Note: These word clouds describe the frequency distribution of words used in the topic that is most strongly negatively correlated with patenting ("Underperforming Benchmark Topic," Topic 14), and the topic that bears the second strongest correlation with patenting ("Operating Performance Topic," Topic 15). As with the innovation topic, these topics are computed from the output of an Latent Dirchlet Allocation (LDA) model fit to a corpus of analyst reports for S&P500 firms. We set the number of topics in the fitted LDA model to be 15.

(a) Underperforming Benchmark Topic ($t = -8.678$)

(b) Operating Performance Topic ($t = 8.697$)

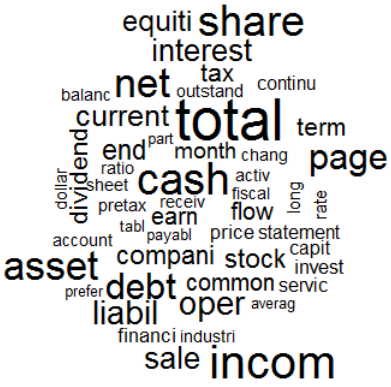


Table A.3: Text-Based Innovation Measure: Word List

Note: This word list describes the frequency distribution of words used in the 'innovation' topic, the top 15 most common words from the topic are listed. The topic itself is from the output of an Latent Dirchlet Allocation (LDA) model fit to a corpus of analyst reports for S&P500 firms. We set the number of topics in the fitted LDA model to be 15, then selected the topic (out of these 15) for which the topic word distribution had the smallest Kullback-Liebler divergence with a benchmark innovation textbook (Tidd, Bessant, and Pavitt, 2005).

Word	Proportion
revenu	0.025
market	0.013
compani	0.012
servic	0.012
growth	0.011
technolog	0.009
product	0.009
network	0.009
system	0.008
softwar	0.007
data	0.007
busi	0.006
custom	0.006
wireless	0.006
total	0.006

Table A.4: Validation of the Textbook Benchmark – Stability of the Patenting Topic Rank Over Time

Note: This table presents evidence on the stability of the quality of the innovation textbook benchmark over our sample period. For each five-year window in our sample, we estimate a separate topic model with 15 topics. For each five-year window, we denote the topic whose loadings at the firm-year level have the maximal correlation with patenting activity as the "Patenting Topic." For each LDA fitted topic (15 per year), we compute the correlation between the word frequencies in the topic and the word frequencies in "Managing Innovation" by Tidd, Bessant, and Pavitt (2005). For each year in this exercise, this table reports the rank of the "Patenting Topic" (out of 15) in terms of the correlation of its word usage with the innovation textbook (a correlation rank of 1 means that the "Patenting Topic" has the highest correlation among all fitted topics with the innovation textbook (15 means the lowest correlation).

Year	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	Pre-2005 Avg.
Correlation Rank	3	2	2	2	2	1	1	1	2	1	1	1.64 / 15
Year	2005	2006	2007	2008	2009	2010						Post-2005 Avg.
Correlation Rank	1	1	1	6	3	1						2.17 / 15

Figure A.2: Robustness to Selecting the Innovation Topic – Alternative Innovation Textbook

Note: This figure presents the Kullback-Liebler (KL) divergence of our selected innovation topic and an alternative source textbook on innovation ("Entrepreneurship and Innovation" by Drucker, published in 1985), and compares it to the average KL divergence from the source textbook on innovation across all of the other topics in the 15-topic LDA fit. The bars indicate the mean KL divergence, and the bands provide 95% confidence intervals computed from the 2.5% and 97.5% percentiles of a bootstrapped sampling distribution with 500 replications.

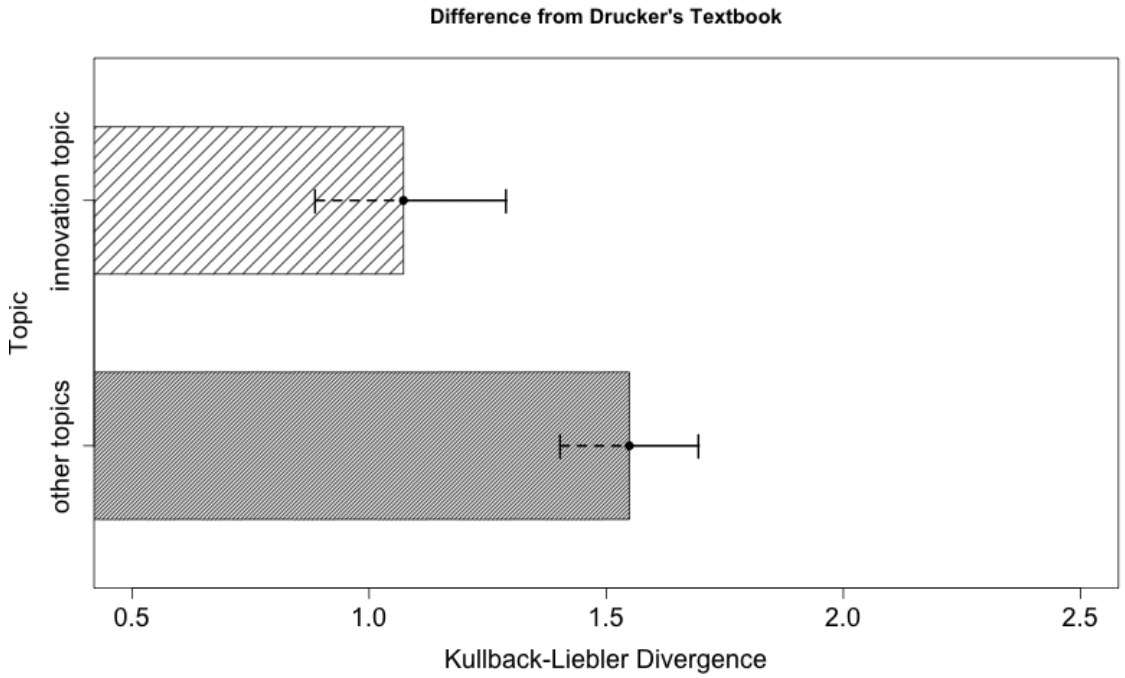
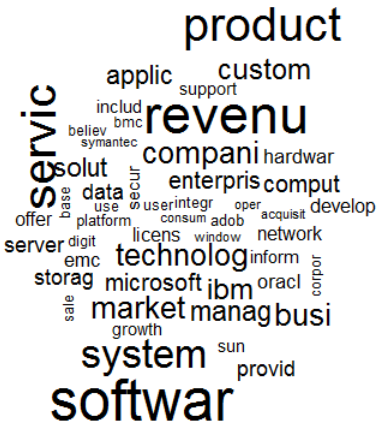


Figure A.3: Word Clouds of Two Innovation Topics from the 50-Topic LDA

Note: These word clouds describes the frequency distribution of words used in the two topics from the 50-topic LDA that are most strongly related to the innovation topic form the 15-topic LDA. As with the innovation topic, these topics are computed from the output of an Latent Dirchlet Allocation (LDA) model fit to the same corpus of analyst reports for S&P500 firms, but this time using 50 topics instead of 15.

(a) First Innovation Topic



(b) Second Innovation Topic

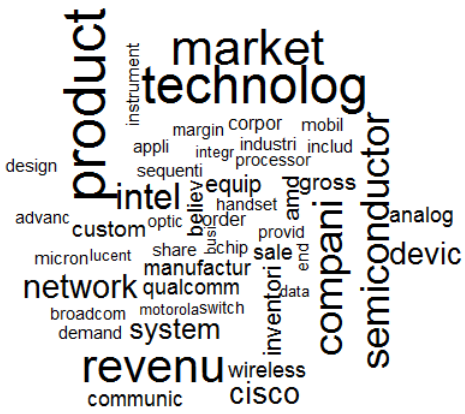


Table A.5: Performance of Firms and Text-Based Innovation (1990-2010) – Using Measure Derived from 50-topic LDA

Note: This table presents OLS regressions that link the text-based innovation measure to measures of performance: ROA, log(Q), and sales growth. In this table, the text-based measure is constructed from a 50-topic LDA instead of a 15-topic LDA. Two topics in the 50-topic LDA are similar to the original innovation topic. We select the topic illustrated in the left panel of Figure A.3 (Topic 6 of 50) because the word frequencies are more strongly correlated with the innovation topic from the 15-topic LDA. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Controls for other innovation measures and firm characteristics are identical to the main specification, but not reported for brevity of presentation. Variable definitions are presented in Table A.1. Standard errors that are clustered by firm are reported in parentheses. Stars *, **,and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

	Return on Assets$_{t+1}$		Log(Q)$_{t+1}$		Sales Growth$_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Based Innovation (Z)$_t$	0.003**	0.003*	0.028***	0.029***	0.009**	0.008
	(0.001)	(0.001)	(0.006)	(0.007)	(0.004)	(0.005)
$\times$ Non-Patenting Firm		0.001		−0.006		0.005
		(0.002)		(0.014)		(0.011)
Controls, Firm FE, SIC2-Year FE	X	X	X	X	X	X
Observations	6,064	6,064	5,931	5,931	6,068	6,068
Adjusted R^2	0.724	0.724	0.799	0.799	0.225	0.224

A.3 Full Results and Alternative Specifications for Performance Regressions

Table A.6: Full Results on Performance of Innovative Firms (1990-2010)

Note: Return on assets is EBITDA scaled by total assets. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. Sales growth is defined as the percentage growth in sales between year t and year t+1 (in decimal form). The text-based innovation measure is converted to a Z-score for ease of interpretability. Patents is the count of granted patents which were applied for during the year. Citations are the forward citations of the patents applied for during the year. Asset tangibility is the property plant and equipment to total assets ratio. Leverage is calculated as the total liabilities over assets with book equity replaced with market equity. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. The market value is as of the last day of the fiscal year. Age is the number of years since the firm entered compustat (with the earliest date 1975). Standard errors that are clustered by firm are reported in parentheses. Stars *, **, and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

(a) Firm Performance

	Dependent variable:								
	ROA_{t+1}			$Log(Q)_{t+1}$			$Salesgrowth_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Text-Innovation $(Z)_t$	0.009^{***}	0.005^{***}	0.005^{***}	0.083^{***}	0.049^{***}	0.039^{***}	0.015^{***}	0.010^{**}	0.013^{**}
	(0.002)	(0.001)	(0.002)	(0.009)	(0.007)	(0.007)	(0.004)	(0.004)	(0.005)
$Log(Patents)_t$	0.002	−0.002	−0.0001	0.003	−0.027	−0.009	−0.007	-0.015^{**}	-0.016^{*}
	(0.003)	(0.003)	(0.003)	(0.015)	(0.016)	(0.016)	(0.005)	(0.007)	(0.008)
$Log(Citations)_t$	0.001	−0.0004	−0.001	0.016^{*}	0.020^{**}	0.007	−0.003	0.002	0.004
	(0.002)	(0.002)	(0.002)	(0.008)	(0.008)	(0.009)	(0.004)	(0.004)	(0.005)
R&D/Assets $(Z)_t$	0.006	0.010^{***}	0.008^{**}	0.074^{***}	0.027	0.023	−0.001	−0.007	−0.010
	(0.005)	(0.004)	(0.004)	(0.020)	(0.022)	(0.023)	(0.006)	(0.009)	(0.010)
$Log(Assets)_t$	−0.001	-0.027^{***}	-0.030^{***}	-0.030^{**}	-0.208^{***}	-0.218^{***}	0.002	-0.071^{***}	-0.064^{***}
	(0.003)	(0.005)	(0.006)	(0.014)	(0.022)	(0.028)	(0.004)	(0.012)	(0.015)
Asset $Tangibility_t$	0.103^{***}	0.055^{***}	0.072^{***}	0.171^{*}	−0.048	0.021	−0.060	-0.308^{***}	-0.260^{***}
	(0.017)	(0.021)	(0.023)	(0.095)	(0.114)	(0.139)	(0.037)	(0.082)	(0.096)
$Leverage_t$	−0.008	−0.008	−0.005	−0.126	-0.127^{*}	-0.154^{**}	-0.086^{***}	−0.055	−0.036
	(0.019)	(0.017)	(0.015)	(0.081)	(0.076)	(0.074)	(0.024)	(0.038)	(0.040)
$Log(Age)_t$	0.002	−0.004	−0.020	-0.079^{**}	−0.149	−0.153	-0.025^{**}	−0.022	-0.105^{*}
	(0.007)	(0.019)	(0.022)	(0.033)	(0.107)	(0.123)	(0.011)	(0.053)	(0.059)
$Cash/Assets_t$	0.102^{***}	0.038	0.054^{**}	0.983^{***}	0.389^{***}	0.394^{***}	0.057	0.014	0.043
	(0.030)	(0.028)	(0.026)	(0.130)	(0.106)	(0.114)	(0.044)	(0.061)	(0.070)
Non-Patenting Firm	−0.009			−0.037			0.00000		
	(0.006)			(0.032)			(0.012)		
Industry (SIC4) FE	X			X			X		
Firm FE		X	X		X	X		X	X
Year FE	X	X		X	X		X	X	
SIC2-Year FE			X			X			X
Observations	6,064	6,064	6,064	5,931	5,931	5,931	6,068	6,068	6,068
Adjusted R^2	0.436	0.674	0.725	0.577	0.771	0.800	0.099	0.159	0.225

Table A.6: Full Results on Performance of Innovative Firms (1990-2010), continued

(b) Firm Performance - Patenting Firm Split

	Dependent variable:								
	ROA_{t+1}			$Log(Q)_{t+1}$			$Salesgrowth_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Text-Based Innovation $(Z)_t$	0.009^{***}	0.005^{***}	0.005^{***}	0.085^{***}	0.052^{***}	0.041^{***}	0.015^{***}	0.009^{*}	0.012^{**}
	(0.002)	(0.002)	(0.002)	(0.010)	(0.008)	(0.008)	(0.005)	(0.005)	(0.006)
× Non-Patenting Firm	0.001	0.001	−0.001	−0.007	−0.003	−0.006	0.003	0.011	0.011
	(0.004)	(0.003)	(0.003)	(0.017)	(0.016)	(0.015)	(0.009)	(0.011)	(0.012)
R&D/Assets $(Z)_t$	0.007	0.010^{***}	0.008^{**}	0.081^{***}	0.029	0.023	−0.005	−0.007	−0.010
	(0.004)	(0.004)	(0.004)	(0.019)	(0.023)	(0.023)	(0.006)	(0.009)	(0.010)
$Log(Assets)_t$	0.001	-0.028^{***}	-0.031^{***}	−0.014	-0.208^{***}	-0.219^{***}	-0.006^{*}	-0.076^{***}	-0.069^{***}
	(0.003)	(0.005)	(0.006)	(0.012)	(0.022)	(0.027)	(0.004)	(0.012)	(0.014)
Asset $Tangibility_t$	0.103^{***}	0.054^{**}	0.071^{***}	0.175^{*}	−0.032	0.021	-0.062^{*}	-0.311^{***}	-0.265^{***}
	(0.017)	(0.021)	(0.023)	(0.096)	(0.114)	(0.138)	(0.037)	(0.082)	(0.096)
$Leverage_t$	−0.007	−0.009	−0.005	−0.120	-0.131^{*}	-0.155^{**}	-0.090^{***}	−0.059	−0.038
	(0.020)	(0.017)	(0.015)	(0.081)	(0.076)	(0.074)	(0.024)	(0.039)	(0.040)
$Log(Age)_t$	0.003	−0.007	−0.022	-0.072^{**}	−0.149	−0.151	-0.030^{***}	−0.039	-0.119^{**}
	(0.007)	(0.020)	(0.023)	(0.032)	(0.105)	(0.120)	(0.011)	(0.054)	(0.060)
$Cash/Assets_t$	0.104^{***}	0.037	0.053^{**}	0.992^{***}	0.386^{***}	0.393^{***}	0.051	0.009	0.041
	(0.031)	(0.028)	(0.026)	(0.130)	(0.107)	(0.114)	(0.045)	(0.062)	(0.070)
Non-Patenting Firm	0.011^{**}			0.063^{**}			−0.011		
	(0.006)			(0.031)			(0.012)		
Industry (SIC4) FE	X			X			X		
Firm FE		X	X		X	X		X	X
Year FE	X	X		X	X		X	X	
SIC2-Year FE			X			X			X
Observations	6,064	6,064	6,064	5,931	5,931	5,931	6,068	6,068	6,068
Adjusted R^2	0.435	0.674	0.725	0.574	0.770	0.800	0.096	0.158	0.224

Table A.7: Performance of Firms and Text-Based Innovation (1990-2010) – Continuous Patenting Interaction, Patenting Firms Only

Note: This table presents OLS regressions that link the text-based innovation measure to measures of performance: ROA, log(Q), and sales growth. As an alternative to Panel B of Table 2, these specifications include *Text-Based Innovation,* as well as an interaction between *Text-Based Innovation* and a continuous measure of patenting (logged number of patents, scaled to have a mean of 0 and standard deviation of 1 for interpretability of the interaction). The sample is restricted to patenting firms only, such that the interaction between *Text-Based Innovation* and *Log(Patents)* directly gives a test of significant differences between firms with high patenting intensity (one standard deviation above) versus average patenting intensity. Standard errors that are clustered by firm are reported in parentheses. Stars *, **, and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

	Dependent variable:								
	Return on Assets$_{t+1}$			Log(Q)$_{t+1}$			Sales Growth$_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Text-Based Innovation (Z)$_t$	0.011***	0.004***	0.004**	0.085***	0.048***	0.039***	0.014***	0.008*	0.012**
	(0.002)	(0.002)	(0.002)	(0.011)	(0.008)	(0.008)	(0.005)	(0.005)	(0.006)
Text-Based Innovation (Z)$_t$ × Log(Patents) (Z)$_t$	0.0004	−0.002	−0.002	0.010	0.014*	0.011	−0.003	−0.004	−0.004
	(0.002)	(0.002)	(0.002)	(0.010)	(0.008)	(0.009)	(0.004)	(0.005)	(0.006)
Log(Patents) (Z)$_t$	0.007	0.001	0.004	0.005	−0.079**	−0.033	−0.014	−0.038**	−0.033*
	(0.006)	(0.006)	(0.006)	(0.030)	(0.032)	(0.035)	(0.011)	(0.015)	(0.018)
Log(Citations)$_t$	−0.0003	−0.001	−0.001	0.016**	0.025***	0.010	−0.002	0.004	0.003
	(0.002)	(0.002)	(0.002)	(0.008)	(0.008)	(0.010)	(0.004)	(0.004)	(0.005)
R&D/Assets (Z)$_t$	0.007	0.009**	0.007*	0.078***	0.033	0.026	−0.001	−0.005	−0.009
	(0.005)	(0.004)	(0.004)	(0.021)	(0.023)	(0.024)	(0.006)	(0.009)	(0.010)
Leverage$_t$	−0.017	−0.014	−0.007	−0.199**	−0.163*	−0.172**	−0.100***	−0.056	−0.023
	(0.025)	(0.020)	(0.018)	(0.101)	(0.085)	(0.088)	(0.028)	(0.042)	(0.044)
Log(Assets)$_t$	−0.001	−0.032***	−0.037***	−0.026	−0.208***	−0.235***	0.002	−0.064***	−0.062***
	(0.004)	(0.006)	(0.006)	(0.018)	(0.025)	(0.032)	(0.005)	(0.014)	(0.016)
Log(Age)$_t$	0.008	−0.016	−0.025	−0.070	−0.160	−0.126	−0.034**	−0.025	−0.099
	(0.011)	(0.027)	(0.029)	(0.044)	(0.135)	(0.153)	(0.015)	(0.068)	(0.072)
Cash/Assets$_t$	0.106***	0.046	0.058*	0.963***	0.424***	0.440***	0.077	0.048	0.086
	(0.033)	(0.031)	(0.030)	(0.151)	(0.122)	(0.125)	(0.050)	(0.068)	(0.079)
Asset Tangibility$_t$	0.114***	0.061***	0.074***	0.145	0.026	−0.016	−0.064	−0.303***	−0.280***
	(0.020)	(0.023)	(0.026)	(0.105)	(0.121)	(0.161)	(0.042)	(0.085)	(0.091)
Industry (SIC4) FE	X			X			X		
Firm FE		X	X		X	X		X	X
Year FE	X	X		X	X		X	X	
SIC2-Year FE			X			X			X
Observations	4,676	4,676	4,676	4,574	4,574	4,574	4,679	4,679	4,679
Adjusted R^2	0.442	0.676	0.717	0.596	0.778	0.798	0.099	0.156	0.203

Table A.8: Performance of Firms and Text-Based Innovation (1990-2010) – Main specification without controlling for patenting measures

Note: This table presents OLS regressions that link the text-based innovation measure to measures of performance: ROA, log(Q), and sales growth. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. After dropping patenting measures, other controls include log(assets), asset tangibility, leverage, log(age), R&D intensity, an indicator for patenting firm, and cash/assets. Variable definitions are presented in Table A.1. Standard errors that are clustered by firm are reported in parentheses. Stars *, **, and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

(a) Firm Performance

	Dependent variable:								
	Return on Assets$_{t+1}$			Log(Q)$_{t+1}$			Sales Growth$_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Text-Based Innovation (Z)$_t$	0.009***	0.005***	0.005***	0.083***	0.051***	0.039***	0.016***	0.012***	0.014***
	(0.002)	(0.001)	(0.002)	(0.009)	(0.007)	(0.007)	(0.004)	(0.004)	(0.005)
R&D/Assets (Z)$_t$	0.007	0.010***	0.008**	0.082***	0.029	0.023	−0.005	−0.007	−0.010
	(0.004)	(0.004)	(0.004)	(0.019)	(0.023)	(0.023)	(0.006)	(0.009)	(0.010)
Leverage$_t$	−0.007	−0.009	−0.005	−0.120	−0.131*	−0.155**	−0.090***	−0.058	−0.038
	(0.020)	(0.017)	(0.015)	(0.081)	(0.076)	(0.074)	(0.024)	(0.039)	(0.040)
Log(Assets)$_t$	0.001	−0.028***	−0.031***	−0.014	−0.208***	−0.219***	−0.006*	−0.076***	−0.069***
	(0.003)	(0.005)	(0.006)	(0.012)	(0.022)	(0.027)	(0.004)	(0.012)	(0.014)
Log(Age)$_t$	0.003	−0.007	−0.022	−0.072**	−0.149	−0.152	−0.030***	−0.037	−0.117**
	(0.007)	(0.020)	(0.023)	(0.032)	(0.105)	(0.120)	(0.011)	(0.054)	(0.060)
Cash/Assets$_t$	0.104***	0.037	0.054**	0.993***	0.386***	0.393***	0.051	0.009	0.040
	(0.030)	(0.028)	(0.026)	(0.130)	(0.107)	(0.114)	(0.045)	(0.062)	(0.070)
Asset Tangibility$_t$	0.103***	0.054**	0.071***	0.175*	−0.032	0.021	−0.062*	−0.310***	−0.265***
	(0.017)	(0.021)	(0.023)	(0.096)	(0.114)	(0.138)	(0.037)	(0.082)	(0.096)
Non-Patenting Firm	−0.011**			−0.062**			0.011		
	(0.006)			(0.030)			(0.012)		
Industry (SIC4) FE	X			X			X		
Firm FE		X	X		X	X		X	X
Year FE	X	X		X	X		X	X	
SIC2-Year FE			X			X			X
Observations	6,064	6,064	6,064	5,931	5,931	5,931	6,068	6,068	6,068
Adjusted R^2	0.435	0.674	0.725	0.575	0.770	0.800	0.096	0.158	0.224

Table A.9: Performance of Firms and Text-Based Innovation (1990-2010), Interactive specifications that control for patent counts and patent citations

Note: This table presents OLS regressions that link the text-based innovation measure to measures of performance: ROA, log(Q), and sales growth. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Other innovation measures – log(patents), log(citations), an indicator for patenting firm, R&D intensity – are included in the specification to provide a basis for comparison. Other controls include log(assets), asset tangibility, leverage, log(age), and cash/assets. The specification presented is perfectly analogous to Panel (b) of Table 2, except that patent counts and citaitons are included as controls in this specification. Variable definitions are presented in Table A.1. Standard errors that are clustered by firm are reported in parentheses. Stars *, **, and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

(a) Firm Performance - Patenting Firm Split

	Dependent variable:								
	ROA_{t+1}			$Log(Q)_{t+1}$			$Salesgrowth_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Text-Based Innovation $(Z)_t$	0.009^{***}	0.005^{***}	0.005^{***}	0.084^{***}	0.050^{***}	0.040^{***}	0.015^{***}	0.008^{*}	0.011^{*}
	(0.002)	(0.002)	(0.002)	(0.010)	(0.008)	(0.008)	(0.005)	(0.005)	(0.006)
Text-Based Innovation $(Z)_t \times$ Non-Patenting Firm	0.001	0.001	−0.001	−0.006	−0.004	−0.006	0.002	0.012	0.012
	(0.004)	(0.003)	(0.003)	(0.018)	(0.016)	(0.015)	(0.009)	(0.011)	(0.012)
R&D/Assets $(Z)_t$	0.006	0.010^{***}	0.008^{**}	0.074^{***}	0.027	0.023	−0.001	−0.007	−0.010
	(0.005)	(0.004)	(0.004)	(0.020)	(0.022)	(0.023)	(0.006)	(0.009)	(0.010)
Leverage$_t$	−0.008	−0.008	−0.005	−0.126	-0.127^{*}	-0.154^{**}	-0.086^{***}	−0.056	−0.036
	(0.019)	(0.017)	(0.015)	(0.081)	(0.076)	(0.074)	(0.024)	(0.038)	(0.040)
Log(Assets)$_t$	−0.001	-0.027^{***}	-0.030^{***}	-0.030^{**}	-0.208^{***}	-0.218^{***}	0.002	-0.071^{***}	-0.064^{***}
	(0.003)	(0.005)	(0.006)	(0.014)	(0.022)	(0.028)	(0.004)	(0.012)	(0.015)
Log(Age)$_t$	0.002	−0.004	−0.020	-0.079^{**}	−0.149	−0.152	-0.025^{**}	−0.024	-0.108^{*}
	(0.007)	(0.019)	(0.023)	(0.033)	(0.107)	(0.123)	(0.011)	(0.053)	(0.060)
Cash/Assets$_t$	0.103^{***}	0.038	0.054^{**}	0.983^{***}	0.389^{***}	0.394^{***}	0.057	0.014	0.043
	(0.030)	(0.028)	(0.026)	(0.130)	(0.105)	(0.114)	(0.044)	(0.061)	(0.070)
Asset Tangibility$_t$	0.103^{***}	0.055^{***}	0.072^{***}	0.171^{*}	−0.048	0.021	−0.060	-0.308^{***}	-0.260^{***}
	(0.017)	(0.021)	(0.023)	(0.095)	(0.114)	(0.139)	(0.037)	(0.082)	(0.096)
Non-Patenting Firm	0.009			0.038			−0.0003		
	(0.006)			(0.032)			(0.013)		
Log(Patents)$_t$	0.002	−0.002	−0.0001	0.003	−0.026	−0.009	−0.007	-0.015^{**}	-0.016^{**}
	(0.003)	(0.003)	(0.003)	(0.015)	(0.016)	(0.016)	(0.005)	(0.007)	(0.008)
Log(Citations)$_t$	0.001	−0.0004	−0.001	0.016^{*}	0.020^{**}	0.007	−0.003	0.002	0.004
	(0.002)	(0.002)	(0.002)	(0.008)	(0.008)	(0.009)	(0.004)	(0.004)	(0.005)
Industry (SIC4) FE	X			X			X		
Firm FE		X	X		X	X		X	X
Year FE	X	X		X	X		X	X	
SIC2-Year FE			X			X			X
Observations	6,064	6,064	6,064	5,931	5,931	5,931	6,068	6,068	6,068
Adjusted R^2	0.436	0.674	0.725	0.577	0.771	0.800	0.099	0.159	0.225

Table A.10: Long-Term Tobin's Q, ROA, and Salesgrowth Using the Text-Based Innovation Measure

Note: Return on assets is EBITDA scaled by total assets. The text-based innovation measure is converted to a Z-score for ease of interpretability. All firms that have at least one patent during the sample period (1990-2004) are included in the regression. Patents is the count of granted patents which were applied for during the year. Asset tangibility is the property plant and equipment to total assets ratio. Leverage is calculated as the total liabilities over assets with book equity replaced with market equity. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. The market value is as of the last day of the fiscal year. Age is the number of years since the firm entered compustat (with the earliest date 1975). Standard errors that are clustered by firm are reported in parentheses. Stars *, **, and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

(a) ROA and Q for time horizons between $t+1$ and $t+4$

	Dependent variable:							
	roa_{t+1}	roa_{t+2}	roa_{t+3}	roa_{t+4}	ln_q_{t+1}	ln_q_{t+2}	ln_q_{t+3}	ln_q_{t+4}
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Text-Based Innovation $(Z)_t$	0.009^{***}	0.006^{***}	0.005^{**}	0.004^{**}	0.083^{***}	0.066^{***}	0.063^{***}	0.059^{***}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.009)	(0.009)	(0.009)	(0.009)
Log Patents_t	0.002	0.003	0.003	0.004	0.003	0.005	0.0004	0.005
	(0.003)	(0.003)	(0.003)	(0.003)	(0.015)	(0.015)	(0.014)	(0.015)
Log Citations_t	0.001	0.0003	0.0002	−0.001	0.016^{*}	0.016^{**}	0.017^{**}	0.013
	(0.002)	(0.002)	(0.002)	(0.002)	(0.008)	(0.008)	(0.008)	(0.008)
R&D Intensity_t	0.006	0.006	0.002	−0.001	0.074^{***}	0.083^{***}	0.077^{***}	0.065^{***}
	(0.005)	(0.004)	(0.004)	(0.004)	(0.020)	(0.022)	(0.021)	(0.020)
Non-Patenting Firm	−0.009	−0.008	−0.010	-0.015^{*}	−0.037	−0.052	-0.060^{*}	-0.071^{**}
	(0.006)	(0.006)	(0.006)	(0.009)	(0.032)	(0.033)	(0.033)	(0.034)
Controls, Industry (SIC4) & Year FE	X	X	X	X	X	X	X	X
Observations	6,064	5,944	5,825	5,710	5,931	5,702	5,514	5,342
Adjusted R^2	0.436	0.450	0.443	0.375	0.577	0.568	0.568	0.570

(b) Sales growth for horizons from $t+1$ through $t+4$

	Dependent variable:			
	salesgrowth_{t+1}	salesgrowth_{t+2}	salesgrowth_{t+3}	salesgrowth_{t+4}
	(1)	(2)	(3)	(4)
Text-Based Innovation $(Z)_t$	0.015^{***}	−0.0004	−0.003	−0.003
	(0.004)	(0.004)	(0.004)	(0.004)
Log Patents_t	−0.007	−0.001	−0.001	−0.004
	(0.005)	(0.005)	(0.005)	(0.004)
Log Citations_t	−0.003	−0.005	−0.002	0.001
	(0.004)	(0.004)	(0.004)	(0.003)
R&D Intensity_t	−0.001	−0.004	−0.009	-0.012^{**}
	(0.006)	(0.006)	(0.006)	(0.006)
Non-Patenting Firm	−0.00000	−0.013	−0.025	0.003
	(0.012)	(0.016)	(0.017)	(0.017)
Controls, Industry (SIC4) & Year FE	X	X	X	X
Observations	6,068	5,943	5,823	5,707
Adjusted R^2	0.099	0.089	0.088	0.093

Table A.11: Performance of Firms and Text-Based Innovation (1990-2010), Standard Errors Double Clustered by Firm and Year

Note: This table presents OLS regressions that link the text-based innovation measure to measures of performance: ROA, log(Q), and sales growth. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Other innovation measures – log(patents), log(citations), an indicator for patenting firm, R&D intensity – are included in the specification to provide a basis for comparison. Other controls include log(assets), asset tangibility, leverage, log(age), and cash/assets. Full results are reported in the appendix (Table A.6). Variable definitions are presented in Table A.1. Standard errors that are double clustered by firm and year are reported in parentheses. Stars *, **, and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

(a) Firm Performance

	Dependent variable:								
	ROA_{t+1}			$Log(Q)_{t+1}$			$Salesgrowth_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Text-Innovation $(Z)_t$	0.009^{***}	0.005^{***}	0.005^{***}	0.083^{***}	0.049^{***}	0.039^{***}	0.015^{***}	0.010^{**}	0.013^{***}
	(0.002)	(0.002)	(0.002)	(0.010)	(0.008)	(0.007)	(0.006)	(0.005)	(0.004)
$Log(Patents)_t$	0.002	−0.002	−0.0001	0.003	−0.027	−0.009	−0.007	-0.015^{**}	-0.016^{**}
	(0.003)	(0.003)	(0.003)	(0.015)	(0.022)	(0.022)	(0.005)	(0.007)	(0.006)
$Log(Citations)_t$	0.001	−0.0004	−0.001	0.016^{*}	0.020^{*}	0.007	−0.003	0.002	0.004
	(0.002)	(0.001)	(0.002)	(0.009)	(0.010)	(0.011)	(0.003)	(0.004)	(0.005)
R&D/Assets $(Z)_t$	0.006	0.010^{**}	0.008^{**}	0.074^{***}	0.027	0.023	−0.001	−0.007	−0.010
	(0.005)	(0.004)	(0.004)	(0.021)	(0.024)	(0.024)	(0.005)	(0.009)	(0.010)
Non-Patenting Firm	-0.009^{*}			−0.037			0.00000		
	(0.006)			(0.031)			(0.009)		
Industry (SIC4) FE	X			X			X		
Firm FE		X	X		X	X		X	X
Year FE	X	X		X	X		X	X	
SIC2-Year FE			X			X			X
Observations	6,064	6,064	6,064	5,931	5,931	5,931	6,068	6,068	6,068
Adjusted R^2	0.436	0.674	0.725	0.577	0.771	0.800	0.099	0.159	0.225

(b) Firm Performance - Patenting Firm Split

	Dependent variable:								
	ROA_{t+1}			$Log(Q)_{t+1}$			$Salesgrowth_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Text-Based Innovation $(Z)_t$	0.009^{***}	0.005^{***}	0.005^{**}	0.085^{***}	0.052^{***}	0.041^{***}	0.015^{***}	0.009^{*}	0.012^{***}
	(0.002)	(0.002)	(0.002)	(0.012)	(0.010)	(0.009)	(0.006)	(0.006)	(0.004)
× Non-Patenting Firm	0.001	0.001	−0.001	−0.007	−0.003	−0.006	0.003	0.011	0.011
	(0.004)	(0.004)	(0.004)	(0.016)	(0.015)	(0.016)	(0.007)	(0.010)	(0.011)
R&D/Assets $(Z)_t$	0.007	0.010^{**}	0.008^{**}	0.081^{***}	0.029	0.023	−0.005	−0.007	−0.010
	(0.005)	(0.004)	(0.004)	(0.021)	(0.025)	(0.024)	(0.005)	(0.008)	(0.010)
Non-Patenting Firm	-0.011^{**}			-0.063^{**}			0.011		
	(0.005)			(0.029)			(0.009)		
Industry (SIC4) FE	X			X			X		
Firm FE		X	X		X	X		X	X
Year FE	X	X		X	X		X	X	
SIC2-Year FE			X			X			X
Observations	6,064	6,064	6,064	5,931	5,931	5,931	6,068	6,068	6,068
Adjusted R^2	0.436	0.674	0.725	0.577	0.771	0.800	0.099	0.159	0.224

A.4 Subsamples and Additional Robustness for Innovation Measure

Table A.12: Performance of Firms and Text-Based Innovation (1990-2010) – Restricting to Firm-Years with Sentiment Below the Median

Note: This table presents OLS regressions that link the text-based innovation measure to measures of performance: ROA, log(Q), and sales growth. Relative to the main specification, the specifications in this table are restricted to firm-year observations for which analyst sentiment is below the median. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Other innovation measures – log(patents), log(citations), an indicator for patenting firm, R&D intensity – are included in the specification to provide a basis for comparison. Other controls include log(assets), asset tangibility, leverage, log(age), and cash/assets. Variable definitions are presented in Table A.1. Standard errors that are clustered by firm are reported in parentheses. Stars *, **,and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

	Dependent variable:								
	ROA_{t+1}			$Log(Q)_{t+1}$			$Salesgrowth_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Text-Innovation $(Z)_t$	0.011^{***}	0.006^{***}	0.006^{**}	0.096^{***}	0.060^{***}	0.049^{***}	0.024^{***}	0.016^{**}	0.021^{*}
	(0.003)	(0.002)	(0.003)	(0.011)	(0.010)	(0.011)	(0.006)	(0.008)	(0.012)
$Log(Patents)_t$	0.001	−0.002	0.001	0.002	−0.031	−0.017	−0.008	−0.008	−0.012
	(0.003)	(0.004)	(0.004)	(0.015)	(0.020)	(0.021)	(0.007)	(0.013)	(0.016)
$Log(Citations)_t$	0.002	−0.0005	−0.002	0.021^{**}	0.027^{***}	0.008	−0.002	0.004	0.009
	(0.002)	(0.002)	(0.002)	(0.009)	(0.010)	(0.013)	(0.005)	(0.007)	(0.008)
R&D/Assets $(Z)_t$	0.005	0.010^{**}	0.009^{*}	0.072^{***}	0.011	0.009	0.005	−0.012	−0.014
	(0.005)	(0.005)	(0.005)	(0.024)	(0.025)	(0.028)	(0.007)	(0.010)	(0.010)
Patenting Firm	0.006			0.040			−0.003		
	(0.007)			(0.033)			(0.017)		
Industry (SIC4) FE	X			X			X		
Firm FE		X	X		X	X		X	X
Year FE	X	X		X	X		X	X	
SIC2-Year FE			X			X			X
Observations	2,965	2,965	2,965	2,965	2,965	2,965	2,965	2,965	2,965
Adjusted R^2	0.438	0.676	0.754	0.599	0.796	0.831	0.096	0.133	0.150

Table A.13: Performance of Firms and Text-Based Innovation (1990-2010) – Restricting to Firm-Years with Highly Forward-Looking Analyst Reports

Note: This table presents OLS regressions that link the text-based innovation measure to measures of performance: ROA, log(Q), and sales growth. Relative to the main specification, the specifications in this table are restricted to firm-year observations for which analysts use more forward-looking language (adapted from MS2015 paper) than the median firm-year observation. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Other innovation measures – log(patents), log(citations), an indicator for patenting firm, R&D intensity – are included in the specification to provide a basis for comparison. Other controls include log(assets), asset tangibility, leverage, log(age), and cash/assets. Variable definitions are presented in Table A.1. Standard errors that are clustered by firm are reported in parentheses. Stars *, **,and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

	Dependent variable:								
	ROA_{t+1}			$Log(Q)_{t+1}$			$Salesgrowth_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Text-Innovation $(Z)_t$	0.009^{***}	0.005^{*}	0.005^{*}	0.089^{***}	0.057^{***}	0.057^{***}	0.021^{***}	0.019^{**}	0.019^{**}
	(0.003)	(0.003)	(0.003)	(0.013)	(0.011)	(0.011)	(0.007)	(0.008)	(0.008)
$Log(Patents)_t$	0.001	−0.009	−0.009	−0.007	-0.055^{**}	-0.055^{**}	-0.015^{*}	−0.016	−0.016
	(0.004)	(0.005)	(0.005)	(0.019)	(0.022)	(0.022)	(0.008)	(0.013)	(0.013)
$Log(Citations)_t$	0.002	0.002	0.002	0.026^{**}	0.030^{***}	0.030^{***}	0.002	0.008	0.008
	(0.002)	(0.002)	(0.002)	(0.010)	(0.010)	(0.010)	(0.005)	(0.007)	(0.007)
R&D/Assets $(Z)_t$	0.004	0.008^{**}	0.008^{**}	0.082^{***}	0.028	0.028	0.002	−0.008	−0.008
	(0.006)	(0.004)	(0.004)	(0.025)	(0.025)	(0.025)	(0.007)	(0.012)	(0.012)
Patenting Firm	0.003			0.002			0.007		
	(0.007)			(0.039)			(0.018)		
Industry (SIC4) FE	X			X			X		
Firm FE		X	X		X	X		X	X
Year FE	X	X		X	X		X	X	
SIC2-Year FE			X			X			X
Observations	2,965	2,965	2,965	2,965	2,965	2,965	2,965	2,965	2,965
Adjusted R^2	0.420	0.682	0.740	0.597	0.786	0.820	0.103	0.162	0.195

Table A.14: Performance of Firms and Text-Based Innovation (1990-2010) – Projecting the Text-Based Innovation Measure onto Industry-Rival (SIC4 x year) Measure

Note: This table presents instrumental variable (IV) regressions that link the text-based innovation measure to measures of performance: ROA, log(Q), and sales growth. In the specifications in this table, we instrument for firm-level text-based innovation using industry-level text-based innovation at the SIC4-year level. This specification removes firm-specific strategic disclosure from the measure by projecting the text-based measure on the level of innovation disclosed by industry rivals. Across specifications, the first-stage F-statistics far exceed conventional thresholds. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Other innovation measures – log(patents), log(citations), an indicator for patenting firm, R&D intensity – are included in the specification to provide a basis for comparison. Other controls include log(assets), asset tangibility, leverage, log(age), and cash/assets. Variable definitions are presented in Table A.1. Standard errors that are double clustered on firm and year are reported in parentheses. Stars *, **,and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

	Dependent variable:					
	ROA_{t+1}		$Log(Q)_{t+1}$		$Salesgrowth_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Instrumented Text-Innovation $(Z)_t$	0.005^{*}	0.004^{*}	0.076^{***}	0.063^{***}	0.006	0.003
	(0.003)	(0.003)	(0.014)	(0.013)	(0.009)	(0.008)
$Log(Patents)_t$	0.001	-0.002	0.002	-0.024	-0.008	-0.016^{**}
	(0.003)	(0.003)	(0.015)	(0.021)	(0.005)	(0.008)
$Log(Citations)_t$	0.001	0.0002	0.0164^{*}	0.019^{*}	0.003	0.003
	(0.002)	(0.002)	(0.009)	(0.010)	(0.003)	(0.004)
$R\&D/Assets\ (Z)_t$	0.007	0.010^{**}	0.075^{***}	0.027	0.0004	-0.007
	(0.006)	(0.004)	(0.021)	(0.024)	(0.005)	(0.009)
Patenting Firm	0.009^{*}		0.037		0.001	
	(0.006)		(0.031)		(0.010)	
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	5,930	5,930	5,930	5,930	5,930	5,930
Adjusted R^2	0.430	0.674	0.577	0.771	0.099	0.159

Table A.15: Performance of Firms and Text-Based Innovation (1990-2010) – Interaction with Post-2005 (Publication of Innovation Textbook) Indicator

Note: This table presents OLS regressions that link the text-based innovation measure to measures of performance: ROA, log(Q), and sales growth. The only difference between these specifications and the main specification is that we include an interaction with a post-2005 indicator to account for whether there is a stronger/weaker relation between text-based innovation after the "Managing Innovation" textbook is published in 2005. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Other innovation measures – log(patents), log(citations), an indicator for patenting firm, R&D intensity – are included in the specification, but their estimates are suppressed for clarity of presentation. Additional controls include log(assets), asset tangibility, leverage, log(age), and cash/assets. Variable definitions are presented in Table A.1. Standard errors that are double clustered on firm and year are reported in parentheses. Stars *, **,and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

	Dependent variable:					
	ROA_{t+1}		$Log(Q)_{t+1}$		$Salesgrowth_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation $(Z)_t$	0.008^{***}	0.004^{**}	0.086^{**}	0.051^{***}	0.013^{**}	0.007
	(0.002)	(0.002)	(0.012)	(0.009)	(0.007)	(0.006)
Text-Innovation $(Z)_t \times$ Post 2005	0.005	0.007	−0.030	−0.014	0.009	0.017
	(0.005)	(0.005)	(0.021)	(0.010)	(0.013)	(0.015)
Other Controls	X	X	X	X	X	X
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	5,930	5,930	5,930	5,930	5,930	5,930
Adjusted R^2	0.431	0.674	0.578	0.771	0.099	0.159

Table A.16: Performance of Firms and Text-Based Innovation (1990-2010) – Innovation Measure Purged of "Revenue" and "Growth" Words

Note: This table presents OLS regressions that link the text-based innovation measure to measures of performance: ROA, log(Q), and sales growth. Relative to the main specifications, the innovation measure is constructed on a textual corpus in which words beginning with "gro" and "rev" are removed before performing the LDA topic analysis. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Other innovation measures – log(patents), log(citations), an indicator for patenting firm, R&D intensity – are included in the specification, but their estimates are suppressed for clarity of presentation. Other controls include log(assets), asset tangibility, leverage, log(age), and cash/assets. Variable definitions are presented in Table A.1. Standard errors that are double clustered on firm and year are reported in parentheses. Stars *, **,and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

	Dependent variable:								
	ROA_{t+1}			$Log(Q)_{t+1}$			$Salesgrowth_{t+1}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Purged Text-Innovation $(Z)_t$	0.010^{***}	0.007^{***}	0.007^{**}	0.085^{***}	0.060^{***}	0.032^{**}	0.017^{***}	0.014^{**}	0.022^{*}
	(0.002)	(0.002)	(0.003)	(0.010)	(0.007)	(0.013)	(0.005)	(0.005)	(0.012)
4-digit SIC Dummies	X			X			X		
Firm FE		X	X		X	X		X	X
Year FE	X	X		X	X		X	X	
Industry-Year FE			X			X			X
Observations	5,930	5,930	5,930	5,930	5,930	5,930	5,930	5,930	5,930
Adjusted R^2	0.432	0.675	0.766	0.577	0.772	0.838	0.100	0.160	0.251

A.5 Additional Results on Innovation Outcomes

Table A.17: Text-Based Innovation Versus Other Aspects of Innovation (1990-2010) – Forecasting Patent Value and Product Announcements

Note: This table presents output from OLS regressions that link our text-based innovation measure to patent counts, citation impact and patenting value. To focus on the within-patenting properties of the innovation measure, the sample is restricted to patenting firms. As in panel (b) of the main table, the dependent variable is the Kogan et al. (2017) measure of market value of patents (i.e., the stock market jump on the day of the granted patent in $millions) aggregated over all patents granted during the year in columns 1 through 4. The dependent variable in columns 5 through 8 is the log of the number of product announcements when the stock market return was above the 75th percentile from Mukherjee, Singh, and Zaldokas (2016). Controls include other innovation measures – log(patents), log(citations), an indicator for patenting firm, R&D intensity, as well as date t values of log(assets), asset tangibility, leverage, log(age), and cash/assets. Standard errors that are double clustered on firm and year are reported in parentheses. Stars *, **, and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

	Dependent variable:							
	Log(1 + Patent Value)$_{t+1}$				Log(1 + Products)$_{t+1}$			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Text Innovation$_t$	0.278***	0.167***	0.085**	0.128***	0.100***	0.046*	0.036	0.044
	(0.059)	(0.038)	(0.040)	(0.038)	(0.031)	(0.027)	(0.028)	(0.029)
Log(1 + Patents)$_t$		0.762***		0.556***		0.034		0.028
		(0.037)		(0.018)		(0.022)		(0.029)
Log(1 + Citations)$_t$		0.138***		0.058***		−0.003		−0.004
		(0.019)		(0.018)		(0.008)		(0.006)
Other Controls		X		X		X		X
Industry (SIC4) FE	X	X			X	X		
Firm FE			X	X			X	X
Year FE	X	X	X	X	X	X	X	X
Observations	3,249	3,249	3,249	3,249	1,617	1,617	1,617	1,617
Adjusted R^2	0.498	0.787	0.815	0.846	0.453	0.546	0.631	0.641

Table A.18: Text-Based Innovation Versus Other Aspects of Innovation (1990-2010) – Negative Text-Based Innovation Measure

Note: This table presents output from OLS regressions that link our "negative" text-based innovation measure to patent counts, citation impact and patenting value. To focus on the within-patenting properties of the innovation measure, the sample is restricted to patenting firms. In panel (a), the dependent variables we consider are logged patent counts over the following three years ($t+1$ to $t+3$), $Log\,(1+Patents_{t+1\to t+3})$ and logged citation impact of patents over the following three years, $Log\left(1+\frac{Citations_{t+1\to t+3}}{Patents_{t+1\to t+3}}\right)$. In panel (b), the dependent variable is the Kogan et al. (2017) measure of market value of patents (i.e., the stock market jump on the day of the granted patent in $millions) aggregated over all patents granted during the year in columns 1 through 4. The dependent variable in columns 5 through 8 is the log of the number of product announcements when the stock market return was above the 75th percentile from Mukherjee, Singh, and Zaldokas (2016). Controls include other innovation measures – log(patents), log(citations), an indicator for patenting firm, R&D intensity, as well as date t values of log(assets), asset tangibility, leverage, log(age), and cash/assets. Standard errors that are clustered by firm are reported in parentheses. Stars *, **, and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

(a) Negative Text-Based Innovation, Patents and Citation Impact

	Dependent variable:							
	Log(1 + Patents$_{t+1\to t+3}$)				Log(1 + $\frac{Citations_{t+1\to t+3}}{Patents_{t+1\to t+3}}$)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Neg. Text-Based Innovation (Z)$_t$	0.188***	−0.004	0.054	0.005	0.009	0.010	−0.014	−0.002
	(0.044)	(0.018)	(0.036)	(0.023)	(0.015)	(0.011)	(0.013)	(0.012)
Log(Patents) $_t$		0.631***		0.418***		−0.297***		−0.169***
		(0.047)		(0.064)		(0.032)		(0.037)
Log(Citations)$_t$		0.281***		0.253***		0.285***		0.127***
		(0.032)		(0.038)		(0.028)		(0.022)
Other Controls		X		X		X		X
Industry (SIC4) FE	X	X			X	X		
Firm FE			X	X			X	X
Year FE	X	X			X	X		
SIC2-Year FE			X	X			X	X
Observations	4,781	4,781	4,781	4,781	3,208	3,208	3,208	3,208
Adjusted R^2	0.608	0.883	0.857	0.908	0.797	0.847	0.909	0.918

(b) Negative Text-Based Innovation, Patent Value, and Product Announcements

	Dependent variable:							
	Log(1 + Patent Value)$_t$				Log(1 + Products)$_t$			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Text-Based Innovation (Z)$_t$	0.254***	−0.011	0.083*	0.014	0.069**	0.006	0.034	0.014
	(0.053)	(0.021)	(0.045)	(0.020)	(0.028)	(0.020)	(0.029)	(0.028)
Log(Patents) $_t$		0.409***		0.555***		0.027		−0.033
		(0.066)		(0.095)		(0.031)		(0.041)
Log(Citations)$_t$		0.673***		0.640***		0.011		0.033
		(0.038)		(0.054)		(0.019)		(0.022)
Other Controls		X		X		X		X
Industry (SIC4) FE	X	X			X	X		
Firm FE			X	X			X	X
Year FE	X	X			X	X		
SIC2-Year FE			X	X			X	X
Observations	4,781	4,781	4,781	4,781	1,715	1,715	1,715	1,715
Adjusted R^2	0.576	0.915	0.832	0.945	0.433	0.536	0.657	0.663

Table A.19: Text-Based Innovation and R&D Expenses (1990-2010)

Note: The dependent variable is the ratio of R&D expenses to total assets. The text-based innovation measure is converted to a Z-score for ease of interpretability. Patents is the count of granted patents which were applied for during the year. Asset tangibility is the property plant and equipment to total assets ratio. Leverage is calculated as the total liabilities over assets with book equity replaced with market equity. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. The market value is as of the last day of the fiscal year. Age is the number of years since the firm entered compustat (with the earliest date 1975). Standard errors that are double clustered on firm and year are reported in parentheses. Stars *, **, and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

	Dependent variable:			
	R&D/Assets$_t$		R&D/Assets$_{t+1}$	
	(1)	(2)	(3)	(4)
Text-Based Innovation (Z)$_t$	0.003^{**}	0.002^{**}	0.003^{***}	0.001^{*}
	(0.002)	(0.001)	(0.002)	(0.001)
Log(Patents)$_t$		0.003^{***}		0.001
		(0.001)		(0.0004)
Patenting Firm		0.004^{**}		0.001^{*}
		(0.002)		(0.001)
Log(Assets)$_t$		-0.005^{***}		-0.001^{**}
		(0.001)		(0.001)
Return on Assets$_t$		-0.020		-0.003
		(0.020)		(0.008)
Asset Tangibility$_t$		0.001		-0.005
		(0.008)		(0.004)
Leverage$_t$		0.005		-0.004
		(0.005)		(0.004)
Log(Age)$_t$		-0.002		-0.002
		(0.003)		(0.001)
R&D/Assets$_t$				0.680^{***}
				(0.083)
Log(Q)$_t$		0.010^{***}		0.003^{**}
		(0.003)		(0.001)
Industry (SIC4) FE	X	X	X	X
Year FE	X	X	X	X
Observations	6,200	6,200	6,074	6,074
Adjusted R^2	0.693	0.713	0.670	0.817

Table A.20: Text-Based Innovation and Lagged Patenting, R&D and Firm Performance (1990-2010)

Note: This table presents OLS regressions that of the text-based innovation measure on lagged measures of performance (ROA), patenting and R&D activity. This table presents the full results from Table 3, which focused on the relation to performance. For ease of interpretation, we standardize the text-based measure to have a mean of 0 and a standard deviation of 1. Other controls include log(assets), asset tangibility, leverage, log(age), cash/assets and an indicator for whether the firm is a patenting firm. Variable definitions are presented in Table A.1. Standard errors that are clustered by firm are reported in parentheses. Stars *, **,and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

	Dependent variable: Text-Based Innovation$_t$				
	(1)	(2)	(3)	(4)	(5)
ROA$_{t-1}$	1.219***	1.120***	0.656**	0.521	0.598
	(0.322)	(0.376)	(0.310)	(0.322)	(0.366)
ROA$_{t-2}$	0.126	0.068	−0.030	0.011	0.039
	(0.316)	(0.397)	(0.327)	(0.332)	(0.409)
ROA$_{t-3}$	0.144	0.133	−0.008	0.013	−0.042
	(0.362)	(0.415)	(0.406)	(0.405)	(0.446)
ROA$_{t-4}$	−0.190	−0.276	−0.484	−0.499	−0.655**
	(0.300)	(0.329)	(0.314)	(0.308)	(0.322)
Log Patents$_{t-1}$	0.024	0.002	−0.056**	−0.050*	−0.015
	(0.026)	(0.027)	(0.026)	(0.027)	(0.031)
Log Patents$_{t-2}$	−0.013	−0.012	−0.047*	−0.040	−0.034
	(0.024)	(0.040)	(0.026)	(0.026)	(0.031)
Log Patents$_{t-3}$	−0.019	−0.025	−0.055**	−0.056**	−0.051
	(0.024)	(0.026)	(0.027)	(0.026)	(0.032)
Log Patents$_{t-4}$	0.014	0.008	−0.052**	−0.042	−0.001
	(0.024)	(0.028)	(0.026)	(0.026)	(0.029)
R&D (Z)$_{t-1}$	−1.389	−0.724	−1.096	−0.470	−0.339
	(1.173)	(1.751)	(1.028)	(1.002)	(1.011)
R&D (Z)$_{t-2}$	2.618***	2.614***	2.264**	2.320***	1.856**
	(0.761)	(0.517)	(0.916)	(0.897)	(0.841)
R&D (Z) $_{t-3}$	1.014	1.132	0.872	0.561	−0.085
	(0.869)	(0.891)	(1.466)	(1.403)	(1.357)
R&D (Z)$_{t-4}$	0.469	0.169	−0.111	−0.076	−0.709
	(0.697)	(0.941)	(0.870)	(0.819)	(0.765)
Other controls		X		X	X
Industry (SIC4) FE	X	X			
Firm FE			X	X	X
Year FE	X	X	X	X	
SIC2-Year FE					X
Observations	3,621	3,621	3,621	3,621	3,621
Adjusted R^2	0.469	0.480	0.557	0.563	0.610

B Building the Text-Based Innovation Measure: Alternative Approaches

B.1 Alternative Scaling of the Topic Loadings in Building the Text-Based Innovation Measure

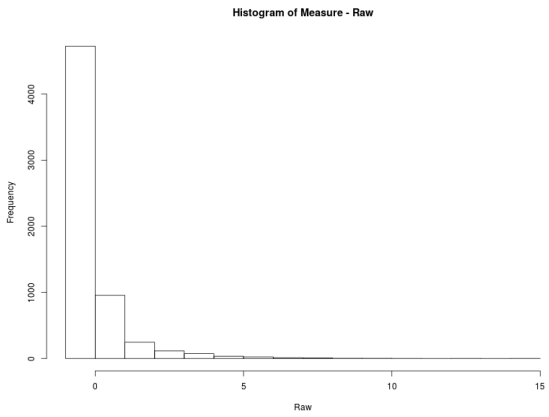
This appendix presents some additional detail on the scaling of the innovation topic loadings underlying the text-based innovation measure. Our primary measure transforms the topic loadings by taking the fourth root before applying the Loughran and McDonald (2011) sentiment filter. We use the fourth root transformation to mitigate skew in the text-based innovation measure.

To highlight the effect of the fourth root transformation, Figure A.4 presents histograms of the topic loadings across analyst reports for the raw topic loadings, an inverse hyperbolic sine transformation (IHS, approximately log), and the fourth root transformation. Both the raw topic loadings and the IHS transformation yield a measure that is highly skewed, whereas the fourth root transformation mitigates the skew, and accordingly should produce a measure with better properties. On this basis, we construct the text-based innovation measure using the less-skewed fourth root transformation.

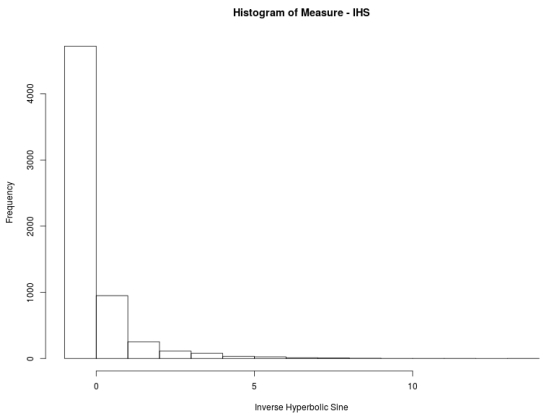
Figure A.4: Histograms of Innovation Topic Loadings and Transformations

Note: This figure presents sample histograms of the topic loadings used as a basis for the text-based innovation measure. In panel (a), the measure is the raw mean of the innovation topic loading for positive analyst reports about the firm over the fiscal year. Panel (b) uses the inverse hyperbolic sine transformation of the raw measure. Panel (c) uses the fourth root of the raw measure, as in the main measure.

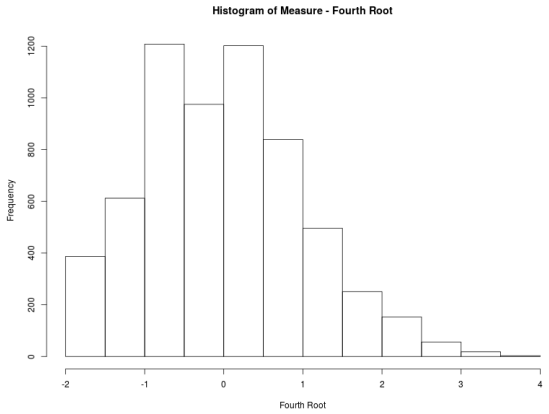
(a) Raw Topic Loadings



(b) Inverse Hyperbolic Sine of Topic Loadings



(c) Fourth Root of Topic Loadings



As a robustness check, we also recreate the measure based on the underlying skewed distributions (using both raw loadings, and IHS transformed loadings) and use these alternative measures in our main specifications. Table A.21 presents the main results using these alternative measures. The sign and significance of the main results are broadly consistent with our main measure, but the estimates tend to be more precise and stable using our primary measure, which confirms the rationale for using the less skewed measure in the first place.

Table A.21: Results Using Alternative Scaling of the Topic Loading to Construct the Text-Based Innovation Measure

Note: Return on assets is EBITDA scaled by total assets. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. SG is sales growth and is defined as the percentage growth in sales between year t and year t+1 (in decimal form). PV is the log of patent value, which it is defined as the stock market jump on the day of the granted patent (in millions) aggregated over all patents granted during the year (see Kogan et al. (2017) for details). V/P is value per patent which is computed as the log of patent value divided by number of patents plus one. FCite/FPat is future citations over future patents, it is computed as the log of the ratio between the number of cites and the number of patents granted in the next three years. The text-based innovation measure is the mean of the innovation topic loading for positive analyst reports about the firm over the fiscal year. In panel (b), the inverse hyperbolic sine of the measure is used. Errors are double clustered on firm and year. Stars *, **, and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

(a) Main Results Using Raw Topic Loadings

	Dependent variable:					
	ROA_{t+1}	$Log(Q)_{t+1}$	SG_{t+1}	PV_{t+1}	V/P_{t+1}	FCite/FPat
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation (raw) $(Z)_t$	0.006^{**}	0.072^{***}	0.009	0.053^{*}	0.062^{**}	0.032
	(0.003)	(0.013)	(0.007)	(0.029)	(0.024)	(0.023)
Other Controls	X	X	X	X	X	X
4-digit SIC Dummies	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Observations	6,064	5,931	6,068	3,249	2,998	3,208
Adjusted R^2	0.432	0.574	0.098	0.805	0.712	0.667
Note:						*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:					
	ROA_{t+1}	$Log(Q)_{t+1}$	SG_{t+1}	PV_{t+1}	V/P_{t+1}	FCite/FPat
	(1)	(2)	(3)	(4)	(5)	(6)
Text-Innovation IHS $(Z)_t$	0.006^{**}	0.072^{***}	0.009	0.054^{*}	0.063^{**}	0.033
	(0.003)	(0.013)	(0.007)	(0.029)	(0.025)	(0.023)
Other Controls	X	X	X	X	X	X
4-digit SIC Dummies	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Observations	6,064	5,931	6,068	3,249	2,998	3,208
Adjusted R^2	0.432	0.574	0.098	0.805	0.712	0.667
Note:						*p<0.1; **p<0.05; ***p<0.01

(b) Main Results Using Inverse Hyperbolic Sine of Topic Loadings

B.2 Word-List Measure versus Latent Dirchlet Allocation

An alternative technique for constructing a measure of innovation from text would be to create a word-list of words related to the idea of innovation. Using a word list of "innovation words," we could measure innovation in one of several ways, for example by counting the number of "innovative words" in each document scaled by the length of the document. As we will see, such an approach — though intuitive — suffers from a number of important limitations . Within the word-list paradigm of textual analysis, there are techniques to overcome these limitations, but these techniques lead to an increase in complexity, and an unsatisfactory level of researcher subjectivity. Our LDA-based method addresses these limitations in a different way, which allows us to avoid any influence of subjectivity on the part of the researcher. In this section, we build the simple word-list measure from the text of analyst reports, and by comparison, highlight some of the strengths of the LDA approach versus an augmented word-list approach..

The first challenge facing word-list approaches is to identify an appropriate list of words for the innovative word-list. Rather than hand classify words that are innovative versus not, we create an objective list by using Princeton University's WordNet database. WordNet is a lexical database available from Princeton University in which nouns, verbs, adjectives, and adverbs are grouped into so called synsets. Each synset contains a set of words with the same distinct meaning (a word is a member of multiple synsets if it has several distinct meanings). A synset represents a unique 'concept'. The database is built as a hierarchy where specific concepts are grouped under more general concepts. For example, rabbit would be grouped under mammals, which are grouped under animals, etc up to the root node 'entity' for all nouns. This type of relation is called hyponomy (or is-a relation, since a rabbit 'is a' mammal), and is the most commonly encoded relation in the WordNet database.[27] We filter out adjectives and adverbs for simplicity of the word-list construction.

[27] Verbs are also grouped into hierarchies, such as hierarchies where the meaning gets more specific (in some sense) further down the tree. Verbs with opposite meaning are linked. In addition to hyponomy, the meronomy relation between nouns is classified, i.e. a part-whole relation

To construct a list of "innovation words," we compute the relatedness between 'innovation' or 'innovate' and all other words in the WordNet database (the two are computed separately),[28] and restrict attention to the top 1% words of most related words. Specifically, we use the Jiang and Conrath (1997) distance to calculate how related two synsets are with each other. To obtain the Jiang and Conrath (1997) distance between two synsets, we compute the sum of all vertices between two synsets in the hierarchy, scaled by their information content. This is calculated as using the least common subsumer, the least general concept that encompasses both synsets. The formula is $JC_D = IC(a) + IC(b) - 2IC(lcs)$, where a and b denote the two synsets. The inverse of the distance is used as the relatedness measure.

Many words have multiple synsets, which indicates that these words have multiple meanings depending on context (e.g., "case" can mean "a small container," "to examine or check out," or "an instance or occurance"). Such words lead to noise in classifying whether words are truly corresponding to their innovative meaning, a problem that we do not have with the LDA-method, which groups words automatically depending on the context that is inferred from the structure of the document. In constructing the word-list measure, we partially address the multiple-meaning problem by using the highest relatedness score to capture the word most closely associated with innovation, but even this solution introduces noise to the extent that analysts are not always using words to mean their most innovative meaning.

We take the resulting word list and measure its similarity with each of our analyst reports by counting how many innovation words each document contains and scaling it by the document length.[29] For consistency with our main LDA-based measure, we aggregate the word-list measure across analyst reports written about the same firm in the same fiscal year for positive reports only (sentiment above the 75th percentile). Tables A.21 and A.22 respectivley present the results performance regressions and patenting regressions that are setup analogously to the tests in the paper. Following the analysis in the main text, we estimate following specifications:,

$$Performance_{it+1} = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \varepsilon_{it} \tag{5}$$

and

$$Patenting_{it+1} = \gamma_t + \xi_s + \beta_1 innov\_text_{it} + \mathbf{X}'_{it}\Gamma + \varepsilon_{it} \tag{6}$$

where $Performance_{ti+1}$ is one of operating performance, log of Q, or salesgrowth; and $Patenting_{ti+1}$ is one of $Log(1 + PatentValue_{ti+1})$, $Log(1 + ValuePerPatent_{ti+1})$, or the log of the ratio of citations to patents over the next three years.

Results in Table A.21 show that this word-list based measure predicts future performance in a way that is quite similar to our LDA-based measure, both in terms of significance and magnitudes, which

[28] The synset for 'innovation' is defined as 'a creation (a new device or process) resulting from study and experimentation'. The synset for 'innovate' is defined as 'bring something new to an environment'.

[29] A popular alternative is to use cosine similarity as in Hoberg and Phillips (2016).

is consistent with how we think of innovation. Nevertheless, the word-list measure fails to correlate in a meaningful manner with more direct measures of innovation. For example, Table A.22 shows that the simplistic word-list measure fails to capture the value of patented innovation, and thus fails our tests that are designed to check whether valuable patented innovation is predicted by the measure of innovation.

It is plausible that the noise introduced by words with multiple meanings leads to enough noise that the word-list measure does not significantly predict the relevant patenting measures. Indeed, the coefficient estimates are of the same sign, just smaller in magnitude and less precisely estimated, by comparison to our LDA-based measure. In this case, refinements of the word-list measure could enhance precision on this dimension. In this spirit, one potential refinement of the word-list measure is called word-sense disambiguation, which is an algorithm aimed at finding the correct meaning of a word in a text. Using a limited sample of analyst reports and firms, we have used a simple Lesk algorithm in this spirit, and though it appears to work well, there is no compelling reason to use an augmented word-list algorithm in this vein over LDA because the augmented word-list algorithm is just as complex, it takes slightly longer to estimate, and it involves more researcher-directed choices that could ultimately influence the results. By contrast, LDA — though complex to estimate — requires much less researcher-input (only the number of topics is selected by the researcher), leading to a stronger, more objective text-based measure of innovation.

Table A.21: Patent Value, Word-List Measure (1990-2010)

Note: Return on assets is EBITDA scaled by total assets. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. Sales growth is defined as the percentage growth in sales between year t and year t+1 (in decimal form). In these tables, we compute the text-based innovation measure analogously as the mean of the innovation word-list loading for positive analyst reports about the firm over the fiscal year. To be consistent with the main measure, we take the fourth root of this measure and convert it to a Z-score. Patents is the count of granted patents which were applied for during the year. Asset tangibility is the property plant and equipment to total assets ratio. Leverage is calculated as the total liabilities over assets with book equity replaced with market equity. Q is defined as the market value of equity plus total assets minus common equity and deferred taxes all divided by total assets. The market value is as of the last day of the fiscal year. Age is the number of years since the firm entered compustat (with the earliest date 1975). Errors are double clustered on firm and year. Stars *, **, and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

	Dependent variable:					
	ROA_{t+1}		$Log(Q)_{t+1}$		$Salesgrowth_{t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
WordList-Innovation $(Z)_t$	0.011^{***}	0.006^{***}	0.073^{***}	0.040^{***}	0.018^{***}	0.015^{***}
	(0.002)	(0.001)	(0.010)	(0.007)	(0.003)	(0.003)
Patenting Firm	0.009		0.039		−0.003	
	(0.005)		(0.031)		(0.009)	
$Log(Patents)_t$	0.002	−0.003	0.022^{**}	−0.006	-0.012^{***}	-0.013^{**}
	(0.002)	(0.002)	(0.011)	(0.014)	(0.003)	(0.006)
$R\&D/Assets\ (Z)_t$	0.006	0.010^{**}	0.079^{***}	0.030	−0.001	−0.006
	(0.005)	(0.004)	(0.022)	(0.025)	(0.005)	(0.008)
$Log(Assets)_t$	−0.0004	-0.026^{***}	−0.022	-0.202^{***}	0.003	-0.069^{***}
	(0.003)	(0.005)	(0.016)	(0.022)	(0.005)	(0.018)
$Asset\ Tangibility_t$	0.102^{***}	0.055^{**}	0.158^{*}	−0.033	-0.061^{*}	-0.305^{***}
	(0.017)	(0.022)	(0.092)	(0.110)	(0.036)	(0.105)
$Leverage_t$	−0.007	−0.007	−0.132	-0.138^{*}	-0.083^{***}	−0.052
	(0.021)	(0.020)	(0.083)	(0.072)	(0.030)	(0.040)
$Log(Age)_t$	0.002	−0.005	-0.083^{**}	−0.166	-0.024^{**}	−0.024
	(0.007)	(0.018)	(0.032)	(0.115)	(0.010)	(0.051)
$Cash/Assets_t$	0.105^{***}	0.040	0.998^{***}	0.397^{***}	0.061	0.017
	(0.030)	(0.029)	(0.121)	(0.094)	(0.049)	(0.054)
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	6,064	6,064	5,931	5,931	6,068	6,068
Adjusted R^2	0.441	0.676	0.577	0.770	0.102	0.161

Note: *p<0.1; **p<0.05; ***p<0.01

Table A.22: Patent Value, Word-List Measure (1990-2010)

Note: The dependent variable is a patent value measure. The first four columns aggregate the value of all patents granted during the year, scaled by patent count in columns 3 and 4. Columns 5 and 6 use the citation weighted patents over the next three years as the measure of patent value. We use patent value data from Kogan et al. (2017) calculated as the abnormal stock market jump (in millions of dollars) on the day of a granted patent. We aggregate these patent values over the fiscal year. In these tables, we compute the text-based innovation measure analogously as the mean of the innovation word-list loading for positive analyst reports about the firm over the fiscal year. To be consistent with the main measure, we take the fourth root of this measure and convert it to a Z-score. Patents is the count of granted patents which were applied for during the year. Other controls are R&D intensity, leverage, the log of total assets, the log of age, and the log of Q. Errors are double clustered on firm and year. Stars *, **, and *** indicate statistical significance at the 10%, 5% and 1% level respectively.

	Dependent variable:					
	$\text{Log}(1+\text{Patent Value})_t$		$\text{Log}(1+\text{Value per Patent})_t$		$\text{Log}(1+\frac{\sum_{s=1}^{3}\text{Citations}_{t+s}}{\sum_{s=1}^{3}\text{Patents}_{t+s}})$	
	(1)	(2)	(3)	(4)	(5)	(6)
$\text{WordList-Innovation}_t$	0.017	0.033*	0.035	0.045**	0.0005	0.010
	(0.017)	(0.017)	(0.022)	(0.017)	(0.015)	(0.014)
$\text{Log}(1+\text{Patents})_t$	0.670***	0.746***				
	(0.032)	(0.042)				
$\text{Log}(1+\text{Citations})_t$ (Z)	0.368***	0.186***	0.124***	0.049*		
	(0.050)	(0.051)	(0.036)	(0.026)		
$\text{R\&D/Assets}_t$	0.859	0.555	-1.211^{*}	-0.037	-1.411^{**}	-0.277
	(0.549)	(0.526)	(0.674)	(0.682)	(0.593)	(0.716)
Levarage_t	-0.791^{***}	-0.687^{***}	-0.845^{***}	-0.767^{***}	-0.053	-0.044
	(0.197)	(0.154)	(0.203)	(0.180)	(0.173)	(0.181)
Log(Assets)_t	0.823***	0.740***	0.420***	0.452***	-0.145^{***}	-0.202^{***}
	(0.049)	(0.070)	(0.041)	(0.073)	(0.039)	(0.061)
Log(Age)_t	-0.068	-0.180	-0.217^{**}	-0.660^{*}	-0.351^{***}	-1.032^{***}
	(0.112)	(0.326)	(0.108)	(0.383)	(0.084)	(0.291)
Log(Q)_t	1.011***	0.909***	0.966***	0.931***	0.156***	-0.0005
	(0.069)	(0.089)	(0.064)	(0.072)	(0.054)	(0.068)
4-digit SIC Dummies	X		X		X	
Firm FE		X		X		X
Year FE	X	X	X	X	X	X
Observations	3,587	3,587	2,999	2,999	3,209	3,209
Adjusted R^2	0.888	0.934	0.710	0.837	0.666	0.799
Note:						*p<0.1; **p<0.05; ***p<0.01