Five Essays in Empirical Finance

Thomas Pauls*

Doctoral Thesis

submitted to
Justus-Liebig-University Gießen,
Department of Business Administration and Economics
March 29, 2017

Disputation:
Mai 30, 2017

Supervisors:
Prof. Dr. Andreas Walter
Prof. Dr. Wolfgang Bessler

*Justus-Liebig-University Gießen | Department of Financial Services
Licher Str. 74 | D-35394 Gießen
E-mail: Thomas.Pauls@wirtschaft.uni-giessen.de | Phone: +49-641-99-22525

Contents

List of tables, figures and appendices 1

I. Analyst herding and investor protection: A cross-country study .. I-3

1. Introduction I-5
2. Data set and variables I-8
3. Methodology I-10
4. Empirical results I-12
5. Conclusion I-18
6. References I-19

II. Trust and the supply side of financial advice II-21

1. Introduction II-23
2. Related research and hypothesis development II-25
3. Data II-28
4. Results II-35
5. Discussion II-40
6. Conclusion II-42
7. References II-43
8. Appendix II-45

III. When do households fail to repay their debt? The role of gender and financial literacy III-48

1. Introduction III-50
2. Data and methodology III-51
3. Results III-52
4. Conclusion III-58
5. References III-59
6. Appendix III-60

IV. Content analysis of business-specific text documents: Introducing a German dictionary IV-61

1. Introduction IV-63
2. Literature IV-64
3. The creation of the BPW dictionary IV-68
4. Evaluation IV-71
5. Conclusion IV-82
6. References IV-84
7. Appendix IV-86

V. CEO-Speeches and stock returns V-87

1. Introduction V-89
2. Literature V-92
3. Data and methodology V-97
4. Results V-103
5. Conclusion V-123
6. References V-125
7. Appendix V-128

Affidavit 129

List of tables, figures and appendices

Tables

Table I-1: Descriptive statistics I-10
Table I-2: Herding results for the overall sample and by country I-12
Table I-3: Herding results split by investor protection I-15
Table I-4: Herding results matrix on investor protection..... I-17
Table II-1: Descriptive statistics II-30
Table II-2: Demographic profiles of bank clienteles II-31
Table II-3: Trust determinants II-33
Table II-4: Bank clienteles and trust in financial advice..... II-36
Table II-5: Robustness – Potential endogeneity of bank choice II-39
Table III-1: Descriptive statistics..... III-53
Table III-2: Probit regressions III-56
Table III-3: IV Regressions with generated instruments III-58
Table IV-1: Number of words in wordlists..... IV-71
Table IV-2: Summary statistics of the quarterly and annual reports IV-73
Table IV-3: Most frequent sentimental words: LM and BPW IV-74
Table IV-4: English vs. German textual sentiment: Reports IV-78
Table IV-5: Summary statistics of the CEO speeches IV-78
Table IV-6: English vs. German textual sentiment: CEO speeches IV-79
Table IV-7: Number of words in wordlists..... IV-80
Table IV-8: Most frequent sentimental words: SENTIWS and LIWC..... IV-81
Table IV-9: Correlations among sentiment measures IV-82
Table V-1: Dictionaries for content analysis..... V-96
Table V-2: Descriptive statistics..... V-103
Table V-3: Correlations V-105
Table V-4: Test of differences of cumulative abnormal returns V-108
Table V-5: Determinants of cumulative abnormal returns V-110
Table V-6: Positive textual sentiment and cumulative abnormal returns.... V-112
Table V-7: Determinants of CARs: Different word lists V-113
Table V-8: Model comparison tests V-115
Table V-9: Determinants of CARs, by weighting schemes employed V-116
Table V-10: Weighted CAR regressions with general language dictionaries V-118
Table V-11: Test of differences of cumulative abnormal trading volumes ... V-119

Table V-12: Determinants of cumulative abnormal trading volume V-120
Table V-13: CAV regressions and weighting V-121
Table V-14: CAV regressions and general language dictionaries V-123

Figures

Figure II-1: Trust in financial advice and general trust II-34
Figure III-1: Debt market participation, financial literacy and gender III-54
Figure III-2: Over-indebtedness, financial literacy and gender III-55
Figure III-3: Marginal effects from Probit models in Table III-2 ..... III-57
Figure IV-1: Correlation plots of quarterly and annual reports IV-76
Figure IV-2: Equivalence tests after Blair and Cole (2002) IV-77
Figure V-1: CARs following the AGM by high vs. low NEG_BPW V-106
Figure V-2: CARs following the AGM by high vs. low TONE_BPW V-107

Appendices

Appendix II-1: Variable descriptions II-45
Appendix II-2: Robustness - Correction of standard error estimates II-46
Appendix II-3: Robustness - Multiple imputations via Rubin's rule II-47
Appendix III-1: Variable descriptions III-60
Appendix IV-1: Adjustments to word independency assumption IV-86
Appendix IV-2: English vs. German textual sentiment: No adjustment IV-86
Appendix V-1: Variable descriptions V-128

I. Analyst herding and investor protection: A cross-country study

Co-authors: Alexander G. Kerl
Own share: 90%

This article has been published as:

Kerl, A. G., & Pauls, T. (2014). Analyst herding and investor protection. A cross-country study. *Applied Financial Economics,* 24(8), 533-542.

Analyst herding and investor protection: A cross-country study

Alexander G. Kerl[a] Thomas Pauls[b]

Abstract - Using a multi-national dataset, we investigate the herding behavior of financial analysts. Our results across a range of different countries suggest that analysts consistently deviate from their true forecasts and issue earnings forecasts that are biased by anti-herding. Furthermore, the level of bias (i.e. anti-herding) seems to be systematically higher for forecasts on companies from European countries compared to the US or Japan. We argue that such differences might stem from diverse levels of investor protection and corporate governance as analysts deviate less from true forecasts when the overall information environment is more transparent and company disclosures are of higher quality. Thereby, we proxy investor protection based on the company-level share of institutional ownership as well as on country-level investor protection measures. Our results show that increasing levels of investor protection and corporate governance mitigate the anti-herding behavior. Especially, when companies that are located in high investor protection countries are held by an increasing number of institutional investors, analysts are most reluctant to issue biased forecasts.

Keywords: Corporate governance; analyst herding; investor protection; earnings forecasts

JEL-Codes: G14; G15; G18

[a] Department of Financial Services, University of Gießen, Licher Str. 74, 35394 Gießen, Germany. Alexander.Kerl@wirtschaft.uni-giessen.de.

[b] Department of Financial Services, University of Gießen, Licher Str. 74, 35394 Gießen, Germany. Thomas.Pauls@wirtschaft.uni-giessen.de.

1. Introduction

Financial analysts serve as information intermediaries in financial markets. It is their job to gather and analyze all available information on a company in order to support investors in their investment decisions. Since analyst research generally contains information value (Asquith et al., 2005), it could be shown in various papers (see, e.g., Brown et al., 2014) that investors actually rely on this information. At the same time, numerous studies found analyst forecasts and recommendations to be systematically biased by herding or anti-herding behavior, calling the information value of analyst research into question.

In this context, herding describes the tendency to 'stick to the crowd' even though the analysts' private information would indicate otherwise.[1] Studies by, for example, Trueman (1994) or Hong et al. (2000) suppose reputational and career concerns to explain analyst herding. The authors believe that analysts herd in order to signal a higher forecasting ability to investors and employers. Thereby, Hong et al. (2000) found young and inexperienced analysts to be especially prone to herding as they are punished more harshly for poor forecast performance. The opposite behavior to herding would be anti-herding. This means that analysts overemphasize their analyses and issue forecasts away from the consensus of precedent forecasts by other analysts.[2] Again, the literature suggests that anti-herding might be explained by career concerns as analysts try to 'stand out from the crowd'. The findings of Clement and Tse (2005), for example, indicate that analysts are more likely to anti-herd with a higher general experience and prior accuracy.

However, both types of biases (i.e. herding as well as anti-herding) basically represent situations where forecasts do not represent the analysts' best knowledge and in that, both constrain the analysts' function as information intermediaries. As different kinds of market participants rely on analyst research to conduct investment decisions, herding as well as anti-herding might eventually skew market prices, foster stock market volatility or contribute to the development of market bubbles.[3] This study contributes to the literature by analyzing the (anti-) herding behavior from a cross-country perspective. Although the literature has generally found that the informativeness and accuracy of analyst research differs

[1] See, for example, Trueman (1994), Hong et al. (2000), Clement and Tse (2005), or Jegadeesh and Kim (2010).

[2] See, for example, Zitzewitz (2001), Bernhardt et al. (2006), Chen and Jiang (2006) or Naujoks et al. (2009).

[3] For a more elaborated discussion on the consequences of herding in financial markets see, for example, De Bondt and Forbes (1999) or Bikhchandani and Sharma (2001).

across different countries (see, e.g., Bhat et al., 2006 or Arand et al., 2015), to the best of our knowledge, it has not yet been analyzed if financial analysts' herding behavior differs, depending on the country a certain company is located in. As a second contribution, we provide first evidence that analysts' deviation from their true estimates depends on the prevailing company- and country-level means of investor protection and corporate governance environment.

Reviewing the research that has been done on analyst herding, one has to differentiate among two different types of forecasts: Stock recommendations and earnings forecasts. With respect to stock recommendations, Welch (2000) not only found that the prevailing consensus has a positive influence on recommendation revisions, but also found that such revisions influence the following two revisions made by consequent analysts. Accordingly, Jegadeesh and Kim (2010) found analysts to herd while issuing stock recommendations. With respect to earnings forecasts, Trueman (1994) also found analysts to issue forecasts not in an unbiased manner, but to herd towards the consensus of previously issued forecasts. He explained the herding behavior with analysts' career and reputational concerns. Similarly, subsequent studies by, for example, Hong et al. (2000) or Clement and Tse (2005) found earnings forecasts to be biased by herding. In contrast, more recent studies (see, e.g., Zitzewitz (2001), Bernhardt et al. (2006), Chen and Jiang (2006) or Naujoks et al. (2009)) emphasized that the previous studies' results might suffer from various problems such as correlated information signals, unexpected common shocks and systematic optimism or pessimism.[4] Thus, they adjusted their methodologies to control for these issues. In contrast to the former studies, they uniformly found analysts to anti-herd, meaning to issue earnings forecasts farther from the consensus than their private information would suggest.

The literature indicates that analyst forecast accuracy is strongly determined by a company's investor protection environment. While Byard, et al. (2006) found company-level investor protection to improve analysts' forecast accuracy, Bhat et al. (2006) revealed the importance of country-level investor protection and corporate governance. To explain this finding, Arand et al. (2015) argued that investor protection leads to high-quality corporate disclosures, and thus, to better inputs for analyst research. Even in more detail, Frankel et al. (2006) showed that the informativeness of analyst research and financial statements are complements. If a higher investor protection environment improves the inputs for analyst research and, consequently, analysts' forecast accuracy, one could reasonably assume that investor protection might also effect analysts' herding or anti-herding behavior.

[4] For a detailed discussion of problems arising within former studies on herding, see Bernhardt et al. (2006).

We hypothesize that an increase in disclosure and information quality eases the assessment of the companies' situation and future earnings expectations not only for analysts themselves but, additionally, for all other market participants. As a consequence, biased forecasts might be recognized more easily, and eventually, analysts might feel compelled to issue forecasts closer to their true estimates. The first indication was provided by Naujoks et al. (2009), who show that German analysts deviate less from their own forecasts in case of larger companies. As the German Corporate Governance Codex plays an increasing role for large companies in Germany, the company's size can be seen as proxy for investor protection.[5]

As a more direct company-level investor protection measure, the literature has revealed a company's share of institutional ownership to be associated with the quality of financial reporting. As Yeo et al. (2002) emphasized, this could be due to the fact that large institutional shareholders have an interest in gathering all available information and, hence, monitor the respective company's management. The authors highlighted that institutional shareholders, by exercising voting rights for example, have the necessary control over the management to enforce their interests. Velury and Jenkins (2006) extended the study of Yeo et al. (2002) and found that large institutional owners indeed fulfil a monitoring role and that a larger fraction of institutional ownership leads to an increased quality of reported earnings.[6] Ljungqvist et al. (2007) even linked ownership directly to the quality of analyst reports. They found that the presence of institutional investors provides incentives for analysts to publish unbiased forecasts, as issuing biased research would undermine their reputation with institutional investors. Ultimately, this is due to the fact that institutions are the primary customers of analyst research. Thus, we hypothesize that analysts are less likely to deviate from their true forecasts (in terms of herding or anti-herding) in case of institutional investors performing monitoring efforts.

A company's investor protection environment can be described not only on the company level, but also on the country level. Based on our cross-country sample, we employ four conceptually different measures of country-level investor protection. The measures describe whether a country's legal system depends on code law or common law (La Porta et al., 1998); the efficiency of a country's anti-self-dealing mechanisms (Djankov et al., 2008); a country's ability of legal enforcement (Leuz et al., 2003) and a country's capability to remedy, prevent and

[5] See Commission of the German Corporate Governance Code (2013).

[6] For further literature on the effect of institutional ownership on the quality of reported earnings, see Shleifer and Vishny (1986), Frankel et al. (2006) or Chen et al. (2007).

punish law violations (Jackson & Roe, 2009). Similar to the company-level argument, we hypothesize that analysts should be less likely to herd or anti-herd when country-level investor protection is high.

Our research, therefore, aims to answer two questions: First, does the (anti-) herding behavior of financial analysts differ between companies located in different countries? And second, to what extent do the company- and country-level measures of investor protection influence the (anti-) herding behavior of analysts? For our analysis, we utilize earnings forecasts from 814,088 analyst reports from January 2005 to June 2010 from eight different countries and employ the herding methodology of Bernhardt et al. (2006).

In line with previous literature (see, e.g., Chen & Jiang, 2006), we find analysts to anti-herd when issuing earnings forecasts. Considering the results for each country separately, anti-herding remains prevalent for all countries in our sample. We find the anti-herding bias to be more severe for forecasts on companies from European countries compared to forecasts on companies from Japan or the US. In addition, our results show that the level of forecast bias (in terms of anti-herding) significantly decreases in case of high levels of company-level investor protection and corporate governance (as proxied by high shares of institutional ownership). Similarly, we find a considerably lower level of forecast bias in case of a strong country-level investor protection environment, as measured by four conceptually different country-level proxies. Finally, as a company's investor protection environment can only be comprehensively described by combining both company- and country-level measures, we analyze both effects simultaneously. Our results show that company-level investor protection only lowers the forecasting bias of analyst research in situations where country-level investor protection and governance are yet strong. In contrast, when country-level investor protection is weak, the additional effect of company-level investor protection has no influence. Thus, we conclude that institutional ownership is not a substitute for but conditional on country-level investor protection when it comes to analyst herding/anti-herding behavior.

This research continues as follows. In Section 2, we describe the data set and variables and in Section 3 we explain the methodology by Bernhardt et al. (2006) that we apply. In Section 4 we present our empirical results before we finally draw a conclusion in Section 5.

2. Data set and variables

We collect analyst report data from FactSet. For each report, our data set contains publication date, earnings per share (EPS) forecast, the actual EPS as

reported at the fiscal year's end, the company's International Securities Identification Number (ISIN) and the country of primary listing. Penny stocks as well as companies that are covered by less than four different analysts per year are dropped from our sample. Our final sample consists of 814,088 analyst reports from January 2005 to June 2010. The reports are written by 9,977 analysts on 3,741 companies located in eight different countries (i.e. France, Germany, Italy, Japan, Spain, Switzerland, the UK and the US). These countries account for more than 59% of the world's total market capitalization[7], and thus, represent the most important financial and economic centers worldwide. Furthermore, they embody different regulatory environments and therefore represent a suitable sample for the purpose of our research.

To measure investor protection and the prevailing corporate governance environment, we apply company- and country-level measures. With respect to the company-level, we gather a company's share of institutional ownership (INSTHOLD) from FactSet. In our sample, it ranges between 0.02% and 100% while the average equals 55.84% and the SD 28.53%. With respect to country-level investor protection, we utilize four conceptually different measures. The first measure (COMMON) is a dummy variable describing the country's legal origin. It equals 1 if the country's legal system depends on common law and 0 in case of code law. According to La Porta et al. (1998), common law countries feature stronger investor protection than code law countries. The second measure is the anti-self-dealing index (ASDI) from Djankov et al. (2008), which addresses the country-level protection of minority shareholders against self-dealing by majority shareholders. The third measure (PUBL_ENF) follows Leuz et al. (2003) and represents a proxy for legal enforcement. It represents the average of three variables from La Porta et al. (1998), namely the efficiency of the judicial system, the rule of law and the level of corruption. The final measure (STAFF_ENF) is taken from Jackson and Roe (2009) and represents a measure of a country's capability to remedy, prevent and punish law violations. It is derived by the number of securities regulator's staff members divided by the country's population. For ASDI, PUBL_ENF and STAFF_ENF, a higher value indicates a higher level of investor protection. Generally, we expect the level of forecast bias (i.e. herding/anti-herding) to be lower in environments of high investor protection.

Table I-1 provides summary statistics for the country-level measures of investor protection. It shows that ASDI is highest for the UK (0.95) and the US (0.65) while it is sharply lower for continental European countries (between 0.27 and 0.42). Looking at PUBL_ENF, the highest values can be found for Switzerland

[7] According to the World Bank as per 2012.

(10.0), the US (9.54) and the UK (9.22). The lowest values can be found for the South-European countries Italy (7.07) and Spain (7.14). For STAFF_ENF, the results appear to be quite similar: the US (23.75) and the UK (19.04) have the highest values, while the levels for continental European countries are severely lower (between 4.43 and 8.87). 462,766 of our 814,088 reports (approximately 56.85%) were issued by US analysts. This is similar to several international studies on analysts like, for example, Barniv et al. (2005) and Jegadeesh and Kim (2006).

Table I-1: Descriptive statistics

This table reports summary statistics of different country-level measures of investor protection for the countries in our sample, namely France, Germany, Italy, Japan, Spain, Switzerland, the UK and the US. COMMON is a dummy variable that equals 1 in case of common law origin and 0 in case of code law origin. ASDI represents the anti-self-dealing index by Djankov et al. (2008). PUBL_ENF represents the legal enforcement index by Leuz et al. (2003). STAFF_ENF is a proxy for a country's capability to remedy, prevent and punish law violations by Jackson and Roe (2009). For ASDI, PUBL_ENF and STAFF_ENF, a higher value indicates a higher level of country-level investor protection. N reflects the number of observations.

Country	COMMON	ASDI	PUBL_ENF	STAFF_ENF	N
France	Code	0.38	8.68	5.91	67,497
Germany	Code	0.28	9.05	4.43	64,830
Italy	Code	0.42	7.07	7.25	23,430
Japan	Code	0.50	9.17	4.32	47,879
Spain	Code	0.37	7.14	8.50	23,234
Switzerland	Code	0.27	10.00	8.87	35,978
United Kingdom	Common	0.95	9.22	19.04	88,474
United States	Common	0.65	9.54	23.75	462,766
Mean		0.59	9.26	17.51	
Median		0.65	9.54	23.75	

3. Methodology

For the purpose of this study, we employ the methodology introduced by Bernhardt et al. (2006). Their methodology is designed to be robust against several methodological issues such as correlated information signals and unexpected market wide earnings shocks that were not addressed in the previous literature. Furthermore, it also accounts for the specific time of information arrival in the forecasting cycle since analysts that issue forecasts later in the course of a year regularly base their forecasts on a richer set of information.[8]

Key assumption of the applied methodology is that an analyst should issue unbiased forecasts incorporating all available information. The probability that the forecast undershoots or overshoots the actual earnings should then be exactly

[8] For a more elaborated discussion on the advantages of this methodology, see Bernhardt et al. (2006).

0.5. Moreover, it should be independent of whether the forecast exceeds or falls short of the consensus which is based on earlier forecasts. We estimate the conditional overshooting and undershooting probabilities as follows:

$$\begin{aligned} p_o &= \Pr\left(F_\tau > A_\tau \mid F_\tau > C_\tau\,;\, F_\tau \neq A_\tau\right) = 0.5 \quad \text{and} \\ p_u &= \Pr\left(F_\tau < A_\tau \mid F_\tau < C_\tau\,;\, F_\tau \neq A_\tau\right) = 0.5 \end{aligned} \tag{1}$$

where p_o is the conditional overshooting probability and p_u is the conditional undershooting probability. Furthermore, τ describes a unique identifier for each analyst report in our sample and F_τ describes the analyst's earnings forecast in report τ. A_τ displays the respective company's actual earnings at the end of the report's forecasting period and C_τ describes the company-specific consensus earnings forecast at the time report τ is published. We estimate the prevailing consensus forecast for any report as the mean of all outstanding earnings forecasts on the respective company and forecast horizon. Thereby, as we expect analysts to incorporate other analysts' forecasts with a time lag, we exclude forecasts that were made on the same day as the report under consideration.[9] Furthermore, as we do not expect analysts to include forecasts that are outstanding for a rather long time and can therefore be considered as stale, we exclude forecasts that have been published 90 days before the report under consideration.[10]

In case of herding, the conditional probabilities to overshoot or undershoot the actual earnings will be smaller than 0.5 ($p_o < 0.5$ and $p_u < 0.5$). In case of anti-herding (i.e. the opposite bias), the conditional probabilities to overshoot or undershoot the actual earnings will be greater than 0.5 ($p_o > 0.5$ and $p_u > 0.5$). To measure herding behavior, (Bernhardt et al., 2006) constructed a test statistic 'S' that is defined as the sample average of the conditional overshooting and undershooting probability estimates. It can be interpreted as a measure of how close analysts issue forecasts to their unbiased estimates based on their private information (i.e. the true forecast). If analysts issue their unbiased best estimates, the S-statistic should be equal to 0.5. A value of S which is lower than 0.5 reveals that analysts are not publishing their unbiased best estimates but rather underemphasize their private information and herd towards the consensus forecast.

[9] While Clement and Tse (2005) use a 3-day lag, excluding reports from only the report's publishing date is consistent with Zitzewitz (2001) and Naujoks et al. (2009).

[10] Including only reports of the last 90 days into the consensus is consistent with Clement and Tse (2005) and Naujoks et al. (2009).

Accordingly, a value of S which is larger than 0.5 means that analysts overemphasize their private information and anti-herd away from the consensus forecast.

4. Empirical results

Within our first analysis as presented in Table I-2, we present not only the herding results for the overall sample (Panel A) but also the results regarding country-specific herding effects (Panel B). Table I-2 is therefore organized as follows: Next to the number of observations N, the table presents unconditional overshooting probabilities, conditional over- and undershooting probabilities, the S-statistic as well as 95% confidence intervals and t-statistics.

Table I-2: Herding results for the overall sample and by country

Notes: The columns of this table are organized as follows: N reflects the number of observations. The unconditional overshooting probability presents the frequency analyst forecasts exceed actual earnings. The conditional overshooting (undershooting) probability depicts the frequency analyst forecasts overshoot (undershoot) the actual earnings, conditional on overshooting (undershooting) the consensus forecast. The S-statistic is the sample average of both conditional probabilities. The null hypothesis of unbiased forecasts translates into $S = 0.5$. Values of S less than (greater than) 0.5 indicate herding (anti-herding) behaviour. Lower and upper bounds of 95% confidence intervals as well as t-statistics are also reported. Panel A of this table reports the S-statistic as introduced by Bernhardt et al. (2006) for our whole sample. Panel B reports the S-statistics for subsamples based on each company's country of primary listing.

	N	Unconditional overshooting probability	Conditional overshooting probability	Conditional undershooting probability	S-Statistic	Lower CI	Upper CI	t-Statistic
Panel A: Whole sample								
	814,088	0.467	0.517	0.568	0.543	0.542	0.544	77.22
Panel B: Country-specific herding								
France	67,497	0.541	0.646	0.560	0.603	0.599	0.607	53.57
Germany	64,830	0.528	0.608	0.548	0.578	0.574	0.582	39.68
Italy	23,430	0.534	0.642	0.548	0.595	0.589	0.602	29.17
Japan	47,879	0.546	0.564	0.464	0.514	0.501	0.519	6.17
Spain	23,234	0.472	0.599	0.649	0.624	0.617	0.630	37.64
Switzerland	35,978	0.533	0.626	0.568	0.597	0.592	0.602	36.77
UK	88,474	0.389	0.459	0.696	0.578	0.574	0.581	45.84
USA	462,766	0.446	0.476	0.559	0.517	0.516	0.519	23.62

With regard to the overall herding analysis, results show that analysts unconditionally overshoot the actual earnings only 46.7% of the time. Under the condition that forecasts exceed the consensus, they exceed the actual earnings 51.7% of the time. Under the condition that they fall short of the consensus, they fall short of the actual earnings 56.8% of the time. The results are similar, albeit smaller, compared to the results of Bernhardt et al. (2006). In their US-sample, they find forecasts to exceed earnings unconditionally only 45% of the time and conditionally to exceed (fall short) earnings 55.6% (62.8%) of the time.

Panel A of Table I-2 also presents the S-statistic for the whole sample. The S-statistic equals 0.543, which means that analysts overshoot the actual earnings in the opposite direction of the consensus by a chance of 54.3%. This result indicates that analysts do not publish their unbiased estimates but instead anti-herd. In other words, analysts highlight their own forecasts by overemphasizing them. This is in line with Bernhardt et al. (2006) and Naujoks et al. (2009), who also found strong evidence for anti-herding based on US and German data.

To the best of our knowledge, the herding behavior of analysts - based on a methodology that is robust to various methodological issues (see Bernhardt et al., 2006) - has not been investigated from a cross-country perspective. Panel B of Table I-2 therefore presents results for different subsamples according to a company's country of primary listing. The most prevalent finding is that forecasts on companies from all countries seem to be biased by anti-herding. Results show S-statistics which are significantly above 0.5 for all different countries. However, comparing the results across all subsamples of countries, we find significant differences not only for the S-statistics, but also for the unconditional and conditional overshooting and undershooting probabilities.

For the UK, the US and Spain, the unconditional overshooting probabilities are below 0.5 and the conditional overshooting probabilities are lower than the conditional undershooting probabilities. Following the argumentation of Bernhardt et al. (2006), this indicates pessimism or the prevalence of positive unforeseen earnings shocks. However, this relation does not hold for the remaining countries of our sample. For France, Germany, Italy, Japan and Switzerland, the unconditional overshooting probabilities are above 0.5 and, at the same time, the conditional overshooting probabilities are higher than the conditional undershooting probabilities. Hence, forecasts on companies from these countries seem to be more optimistic, or alternatively, more frequently challenged by negative unforeseen earnings shocks. For Germany, our results are in line with Naujoks et al. (2009), who analyzed forecasts on German companies from 1994 to 2005.

Investigating the herding behavior for analysts forecasting companies from European countries, our results report the highest S-statistics for companies from Spain (0.624) and France (0.603). The anti-herding bias for forecasts on companies from Switzerland and Italy is somewhat smaller with corresponding S-statistics of 0.597 and 0.595, respectively. For forecasts on companies from Germany and the UK, we find the S-statistic to equal 0.578. This is close to the results of Naujoks et al. (2009), who reported the S-statistic for forecasts on German companies to equal 0.583. For the US, our results reveal a much lower degree of anti-herding compared to European countries. In fact, the S-statistic for forecasts on US companies is the second lowest in our sample as the respective S-statistic

equals 0.517. Hence, it seems as if at least for this very recent time period, forecasts on US companies are much less biased compared to forecasts on all other countries including Europe. In contrast, Bernhardt et al. (2006), who also analyzed forecasts on US companies found a much higher S-statistic of 0.592 for the period from 1989 to 2001. However, the anti-herding bias in our more recent sample period seems to be much lower. As our sample ranges from 2005 to 2010, a change in analyst behavior over time might explain the difference to Bernhardt et al. (2006). Finally, forecasts on Japanese companies suffer the lowest levels of anti-herding in our sample with a corresponding S-statistic of 0.514.

Anti-herding and investor protection

Among others, Yeo et al. (2002) and Velury and Jenkins (2006) have shown that the presence of institutional owners is positively associated with the quality of companies' information disclosures. We, therefore, hypothesize that the improved information quality eases the assessment of the companies' situation and future earnings for all kinds of market participants. Consequently, one can assume that biased analyst forecasts would be recognized more easily by other market participants. Hence, in case of high quality disclosures as proxied by the presence of institutional ownership, analysts might feel compelled to issue forecasts that are less biased (by anti-herding effects) and closer to their true estimates.

Within Panel A of Table I-3, we therefore present evidence concerning the herding behavior of analysts, relative to the prevailing level of institutional ownership. Results show that the unconditional overshooting probabilities are above 0.5 in case of low levels of institutional ownership (quintile 1 and 2) and below 0.5 for high levels of institutional ownership (quintile 3 to 5). Similarly, the conditional overshooting probabilities are higher than the conditional undershooting probabilities within the two lowest quintiles of INSTHOLD, whereas this association reverses for higher levels of institutional ownership. Hence, it seems as if analysts are less likely to issue optimistic forecasts along increasing levels of institutional ownership. With respect to the S-statistic that is consistently above 0.5, our results show that forecasts on companies with different levels of institutional ownership are biased by anti-herding behavior. However, within an increasing level of institutional ownership, we find the anti-herding bias to shrink. For the quintile of companies with the lowest levels of institutional ownership, the S-statistic equals 0.574. It decreases to a value of 0.517 for the quintile of companies with the highest levels of institutional ownership. Hence, analysts' forecasts not only are less optimistic in case of a higher share of institutional ownership but also appear to be less biased by anti-herding behavior and are, therefore, closer to the analysts' true estimates. Apart from the company-specific level of investor

protection, one might also proxy protection and corporate governance levels by country-specific measures.

Table I-3: Herding results split by investor protection

Notes: Panel A of this table reports the S-statistics for subsamples based on a company's share of institutional ownership (INSTHOLD). The quintiles are ordered from low (quintile 1) to high (quintile 5) institutional ownership. Panel B shows herding results based on subsamples of low and high investor protection and corporate governance environments. COMMON is a dummy variable that equals 1 in case of common law origin and 0 in case of code law origin. For all other variables, we split the sample into subsamples based on the median. ASDI represents the anti-self-dealing index by Djankov et al. (2008). PUBL_ENF represents the legal enforcement index by Leuz et al. (2003). STAFF_ENF is a proxy for a country's capability to remedy, prevent and punish law violations by Jackson and Roe (2009). The columns of this table are organized as follows: N reflects the number of observations. The unconditional overshooting probability presents the frequency analyst forecasts exceed actual earnings. The conditional overshooting (undershooting) probability depicts the frequency analyst forecasts overshoot (undershoot) the actual earnings, conditional on overshooting (undershooting) the consensus forecast. The S-statistic is the sample average of both conditional probabilities. The null hypothesis of unbiased forecasts translates into $S = 0.5$. Values of S less than (greater than) 0.5 indicate herding (anti-herding) behaviour. Lower and upper bounds of 95% confidence intervals as well as t-statistics are also reported.

Sample	N	Unconditional overshooting probability	Conditional overshooting probability	Conditional undershooting probability	S-statistic	Lower CI	Upper CI	t-Statistic
Panel A: Quintiles of institutional ownership								
1 (low)	162,835	0.517	0.596	0.552	0.574	0.572	0.577	59.78
2	162,849	0.516	0.585	0.541	0.563	0.560	0.565	50.67
3	162,809	0.459	0.506	0.569	0.538	0.535	0.540	30.38
4	162,825	0.421	0.458	0.599	0.528	0.526	0.531	22.76
5(high)	162,770	0.424	0.450	0.583	0.517	0.514	0.519	13.48
Panel B: Country-level measures of investor protection								
By COMMON								
Code	262,848	0.531	0.615	0.546	0.581	0.579	0.583	82.65
Common	551,240	0.437	0.473	0.580	0.526	0.525	0.528	39.14
By ASDI								
low	191,539	0.527	0.624	0.568	0.596	0.594	0.598	83.89
high	622,549	0.449	0.485	0.568	0.527	0.526	0.528	42.37
By STAFF_ENF								
low	203,636	0.537	0.615	0.531	0.573	0.571	0.575	65.94
high	610,452	0.444	0.487	0.582	0.534	0.533	0.536	53.31
By PUBL_ENF								
low	178,991	0.527	0.625	0.565	0.595	0.593	0.598	80.75
high	635,097	0.451	0.488	0.569	0.529	0.527	0.530	45.48

Panel B of Table I-3, therefore, presents results for high versus low investor protection subsamples, as measured by country-level proxies. Apart from a sample-split into common and code law origin, we also split the sample into high and low investor protection subsamples based on the ASDI, the staff enforcement index (STAFF_ENF) and the public enforcement index (PUBL_ENF), as explained in Section 2. The results for the unconditional and conditional overshooting and undershooting probabilities are similar to the results based on using the

company-level investor protection measure. The unconditional overshooting probabilities are above 0.5 for low levels of country-level investor protection, while they shrink to values below 0.5 for high levels. Furthermore, the differences between the conditional overshooting and conditional undershooting probabilities are positive for low country-level investor protection environments and negative for high country-level investor protection environment in terms of all four used measures. Overall, it is unlikely that increasing levels of investor protection and corporate governance lead analysts to become more pessimistic about the respective companies. Therefore, one might follow Bernhardt et al.'s (2006) second explanation for positive differences between conditional over- and undershooting probabilities, which argues that analysts are less often surprised by negative earnings shocks. This seems reasonable as the analysts' potential to identify prospective earnings risks should be fostered by the improved information which we assume to come in hand with a higher investor protection environment. Looking at the S-statistics for the subsamples based on the country-level measures of investor protection, Panel B of Table I-3 proves anti-herding to be severely lower for forecasts on companies from common law countries ($S = 0.526$) compared to those on companies from code law countries ($S = 0.581$). Similar results are found for all other three proxies of country-specific investor protection. In case of high investor protection and corporate governance (i.e. above median levels of ASDI, STAFF_ENF and PUBL_ENF), S-statistics are much lower compared to the respective subsamples of low investor protection environments. Nevertheless, as the S-statistics remain above 0.5, we find anti-herding throughout all of our subsamples although the forecast bias within high investor protection environments is much lower in relative terms.

So far, we have shown that both company- and country-level investor protection measures influence the forecasting behavior of analysts. Within the next analysis, we now combine both effects. Therefore, we provide S-statistics for all kinds of combinations between company- and country-level investor protection levels.

Table I-4: Herding results matrix on investor protection

Notes: This table is organized as follows: each row of the table represents one quintile based on the company-level investor protection (i.e. the share of institutional ownership). The quintiles are ordered from low (quintile 1) to high (quintile 5) share of institutional ownership. With respect to each column, the sample is split based on country-level investor protection (i.e. low and high investor protection and corporate governance environments). For the legal origin, we differentiate between code and common law. COMMON is a dummy variable that equals 1 in case of common law origin and 0 in case of code law origin. For all other variables, we split the sample into subsamples based on the median. ASDI represents the anti-self-dealing index by Djankov et al. (2008). PUBL_ENF represents the legal enforcement index by Leuz et al. (2003). STAFF_ENF is a proxy for a country's capability to remedy, prevent and punish law violations by Jackson and Roe (2009). For each combination of company- and country-level investor protection, we provide the subsample's S-statistic and the number of observations (in parentheses). Q5-Q1 computes the difference in the S-statistic between the highest and the lowest quintile. The reported t-statistic's null hypothesis analyses whether the difference Q5-Q1 is equal to zero.

	Legal Origin		ASDI		STAFF_ENF		PUBL_ENF	
INSTHOLD	Code	Common	low	high	low	high	low	high
	(N)	(N)	(N)	(N)	(N)	(N)	(N)	(N)
1 (low)	0.583	0.537	0.594	0.538	0.573	0.570	0.597	0.543
	(52,682)	(110,258)	(38,360)	(124,515)	(40,730)	(122,137)	(35,802)	(127,025)
2	0.579	0.531	0.601	0.536	0.572	0.536	0.601	0.536
	(52,478)	(110,332)	(38,360)	(124,510)	(40,743)	(122,222)	(35,882)	(127,218)
3	0.572	0.524	0.594	0.525	0.560	0.525	0.591	0.527
	(52,562)	(110,219)	(38,307)	(124,537)	(40,719)	(121,924)	(35,738)	(126,902)
4	0.572	0.533	0.591	0.533	0.570	0.533	0.592	0.532
	(52,673)	(110,345)	(38,239)	(124,600)	(40,766)	(122,136)	(35,785)	(126,995)
5 (high)	0.596	0.508	0.600	0.509	0.591	0.509	0.596	0.510
	(52,453)	(110,086)	(38,273)	(124,387)	(40,678)	(122,033)	(35,784)	(126,957)
Q5-Q1	0.012	-0.029	0.006	-0.03	0.018	-0.061	-0.001	-0.033
t-Statistic	3.96	-13.69	1.60	-14.78	5.16	-30.20	-0.36	-16.78

Table I-4 is organized as follows: each row of the table represents one quintile based on the company-level investor protection (i.e. the share of institutional ownership). The quintiles are ordered from low (quintile 1) to high (quintile 5) share of institutional ownership. With respect to each column, the sample is split based on country-level investor protection (i.e. low versus high investor protection and corporate governance environments). Quite interestingly, all differences in S-statistics between the lowest and highest quintile of institutional ownership (Q5-Q1) for the subsamples of low country-level investor protection appear quite low and are, at least partly, not statistically significant. For code law countries, for example, the S-statistic of the low ownership quintile ($S = 0.583$) almost equals the S-statistic of the high ownership quintile ($S = 0.596$). Similar findings apply to all subsamples of low investor protection. Hence, our results do not reveal any positive influence (i.e. forecast bias decreasing effect) based on the presence of institutional ownership, conditional on low investor protection and governance countries. On the contrary, once we purely focus on countries with high levels of investor protection and strong corporate governance, all differences in S-statistics between the lowest and highest quintile of institutional ownership (Q5-Q1) are substantial and highly significant. For common law countries, for example, the

S-statistic of the low ownership quintile equals 0.537 whereas it sharply decreases to 0.508 for the high ownership quintile. Table I-4 summarizes similar findings for all other subsamples of high country-level investor protection. Overall, the corresponding S-statistics are very close to 0.5. Hence, analysts are very reluctant to issue biased forecasts in situations of strong country-level investor protection, possibly due to an overall increase in information quality. Nevertheless, large shareholders' ability to put pressure on companies' management in order to improve the information quality comes only into effect in high investor protection environments. This indicates that institutional ownership is not able to serve as a substitute for a lack of country-level investor protection with respect to analyst herding. However, if high institutional ownership and a high country-level investor protection environment come in hand, the combination is very effective in bringing analysts to issue forecasts close to the analysts' best estimates.

5. Conclusion

To the best of our knowledge, the herding behavior of analysts has not yet been investigated in a cross-country study that, at the same time, employs a methodology robust to methodological issues like, for example, correlated information signals, unexpected market wide earnings shocks or systematic optimism.

For the whole sample, our results show that analysts anti-herd with respect to their earnings forecasts. Anti-herding represents a situation where analysts overemphasize their private information and, therefore, anti-herd away from the consensus of precedent analysts. Our results are consistent with research on analyst herding by Zitzewitz (2001), Bernhardt et al. (2006), Chen and Jiang (2006) or Naujoks et al. (2009).

However, while using a multi-national dataset, we contribute to the literature by showing that all forecasts are biased by anti-herding, irrespective of the country that we focus on. Thereby, our results show more severe anti-herding behavior for forecasts on companies from European countries compared to forecasts on companies from Japan or the US. In addition, we hypothesize that the cross-country differences stem from the company's investor protection and corporate governance environment. This might be due to the fact that the overall investor protection environment improves a company's disclosure quality which eases the assessment of the company's situation and future earnings for all kinds of market participants. Consequently, biased analyst forecasts could be recognized more easily by other market participants. Hence, analysts might be reluctant to issue biased forecasts and remain closer to their true estimates. Our results back this argumentation since higher company-level investor protection, as proxied by the

share of institutional ownership, significantly reduces the anti-herding behavior of analysts. Similarly, strong country-level investor protection and corporate governance environments, as measured by a country's common versus code law origin, the efficiency of a country's anti-self-dealing mechanisms, a country's ability of legal enforcement and a country's capability to remedy, prevent and punish law violations, also sharply reduce forecast biases of analysts.

Finally, as a company's investor protection environment can only be comprehensively described by combining company-level and country-level means of investor protection, we investigate the combined effect of a company's share of institutional ownership and the different country-level measures of investor protection. We find that institutional ownership cannot serve as a substitute for country-level investor protection when it comes to analyst anti-herding as its effect does only come into play in environments of high investor protection. Consequently, based on our results, analyst forecasts are least biased for companies with high shares of institutional ownership which are located in countries with high investor protection and corporate governance environments.

6. References

Arand, D., Kerl, A. G., & Walter, A. (2015). When do sell-side analyst reports really matter? Shareholder protection, institutional investors and the imformativeness of equity research. *European Financial Management*, 21(3), 524-555.

Asquith, P., Mikhail, M. B., & Au, A. S. (2005). Information content of equity analyst reports. *Journal of Financial Economics*, 75(2), 245-282.

Barniv, R., Myring, M. J., & Thomas, W. B. (2005). The association between the legal and financial reporting environments and forecast performance of individual analysts. *Contemporary Accounting Research*, 22(4), 727-758.

Bernhardt, D., Campello, M., & Kutsoati, E. (2006). Who herds? *Journal of Financial Economics*, 80(3), 657-675.

Bhat, G., Hope, O.-K., & Kang, T. (2006). Does corporate governance transparency affect the accuracy of analyst forecasts? *Accounting & Finance*, 46(5), 715-732.

Bikhchandani, S., & Sharma, S. (2001). Herd behavior in financial markets. *IMF Staff Papers*, 47(3), 279-310.

Brown, N. C., Wei, K. D., & Wermers, R. (2014). Analyst recommendations, mutual fund herding, and overreaction in stock prices. *Management Science*, 60(1), 1-20.

Byard, D., Li, Y., & Weintrop, J. (2006). Corporate governance and the quality of financial analysts' information. *Journal of Accounting and Public Policy*, 25(5), 609-625.

Chen, Q., & Jiang, W. (2006). Analysts' weighting of private and public information. *The Review of Financial Studies*, 19(1), 319-355.

Chen, X., Harford, J., & Li, K. (2007). Monitoring: which institutions matter? *Journal of Financial Economics*, 86(2), 279-305.

Clement, M. B., & Tse, S. Y. (2005). Financial analyst characteristics and herding behavior in forecasting. *The Journal of Finance*, 60(1), 307-341.

Commission of the German Corporate Governance Code. (2013). German Corporate Governance Code. Retrieved from http://www.corporate-governance-code.de/index-e.html

De Bondt, W. F. M., & Forbes, W. P. (1999). Herding in analyst earnings forecasts: evidence from the United Kingdom. *European Financial Management*, 5(2), 143-163.

Djankov, S., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2008). The law and economics of self-dealing. *Journal of Financial Economics*, 88(3), 430-465.

Frankel, R., Kothari, S. P., & Weber, J. (2006). Determinants of the informativeness of analyst research. *Journal of Accounting and Economics*, 41(1-2), 29-54.

Hong, H., Kubik, J. D., & Solomon, A. (2000). Security analysts' career concerns and herding of earnings forecasts. *The RAND Journal of Economics*, 31(1), 121-144.

Jackson, H. E., & Roe, M. J. (2009). Public and private enforcement of securities laws: Resource-based evidence. *Journal of Financial Economics*, 93(2), 207-238.

Jegadeesh, N., & Kim, W. (2006). Value of analyst recommendations: international evidence. *Journal of Financial Markets*, 9(3), 274-309.

Jegadeesh, N., & Kim, W. (2010). Do analysts herd? An analysis of recommendations and market reactions. *The Review of Financial Studies*, 23(2), 901-937.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W. (1998). Law and finance. *Journal of Political Economy*, 106(6), 1113-1155.

Leuz, C., Nanda, D., & Wysocki, P. D. (2003). Earnings management and investor protection: an international comparison. *Journal of Financial Economics*, 69(3), 505-527.

Ljungqvist, A., Marston, F., Starks, L. T., Wei, K. D., & Yan, H. (2007). Conflicts of interest in sell-side research and the moderating role of institutional investors. *Journal of Financial Economics*, 85(2), 420-456.

Naujoks, M., Aretz, K., Kerl, A. G., & Walter, A. (2009). Do German security analysts herd? *Financial Markets and Portfolio Management*, 23(1), 3-29.

Shleifer, A., & Vishny, R. W. (1986). Large shareholders and corporate control. *Journal of Political Economy*, 94(3), 461-488.

Trueman, B. (1994). Analyst forecasts and herding behavior. *The Review of Financial Studies*, 7(1), 97-124.

Velury, U., & Jenkins, D. S. (2006). Institutional ownership and the quality of earnings. *Journal of Business Research*, 59(9), 1043-1051.

Welch, I. (2000). Herding among security analysts. *Journal of Financial Economics*, 58, 369-396.

Yeo, G. H., Tan, P. M., Ho, K. W., & Chen, S. (2002). Corporate Ownership Structure and the Informativeness of Earnings. *Journal of Business Finance & Accounting*, 29(7-8), 1023-1046.

Zitzewitz, E. (2001). Measuring herding and exaggeration by equity analysts and other opinion sellers. *Stanford GSB Working Paper No. 1802.*

II. Trust and the supply side of financial advice

Co-authors: Oscar A. Stolper, Andreas Walter
Own share: 45%

This paper was presented on the following refereed conferences/ workshops:

- PhD Workshop of the German Finance Association (DGF), Leipzig, Germany, 2015.
- World-Finance Conference, New York, USA, 2016.

This paper was presented on the following non-refereed conferences/ workshops:

- International doctoral seminar in banking and finance, Rauischholzhausen, Germany, 2015.

Trust and the supply side of financial advice

Thomas Pauls[a] Oscar A. Stolper[b] Andreas Walter[c]

Abstract - In this study, we investigate how two key dimensions of trust formation, i.e. interpersonal trust in the advisor (narrow-scope trust) and broader trust in the business context in which the advisor operates (broad-scope trust), impact households' overall trust in financial advice. To capture the potential influence of broad-scope trust, we make use of novel survey data obtained from the Panel on Household Finances (PHF) and contrast households' propensity to trust financial advice provided by advisors employed at community banks versus large banks, which have been shown to feature fundamentally different trust profiles. We document that financial advice provided by large-bank advisors is significantly less likely to be trusted, thus rejecting the notion that trust in financial advice is essentially equivalent to trusting one's financial advisor. Instead, we provide strong evidence in support of an integrated conceptualization of clients' trust in financial advice, which highlights the importance of establishing broad-scope trust.

Keywords: Financial advice, trust, household finance, Panel on Household Finances (PHF)

JEL-Codes: D12, D14, G20

[a] Department of Financial Services, University of Gießen, Licher Str. 74, 35394 Gießen, Germany. Thomas.Pauls@wirtschaft.uni-giessen.de.

[b] Institute of Accounting and Finance, University of Marburg, Am Plan 1, 35032 Marburg, Germany. Oscar.Stolper@wiwi.uni-marburg.de.

[c] Department of Financial Services, University of Gießen, Licher Str. 74, 35394 Gießen, Germany. Andreas.Walter@wirtschaft.uni-giessen.de.

1. Introduction

In light of an increasing responsibility of households for the planning of their personal finances along with the well-documented lack of financial literacy[1] to master this task autonomously, seeking expert financial advice seems a beneficial step for consumers to take in order to arrive at informed financial decisions. Indeed, according to a recent poll in Germany, as much as 81% of all households report the financial advisor at their house bank to be the single source of information to consult when it comes to financial matters (DSGV, 2014). Chater et al. (2010) reach similar conclusions in their large-scale survey of advisees across eight member countries of the European Union (EU): 80% of households interacted with a personal advisor prior to purchasing investment products.

Moreover, trustworthiness is the key criterion when selecting a financial advisor (Johnson & Grayson, 2005; Lachance & Tang, 2012) and most advisees indeed have a high level of trust in the advisor with whom they consult (e.g. Mullainathan et al., 2013; Monti et al., 2014; Gennaioli et al., 2015), even though the literature documents a largely negative record of expert advice when it comes to improving households' financial decisions (e.g. Bergstresser et al., 2009; Bhattacharya et al., 2012; Mullainathan et al., 2013; Von Gaudecker, 2015). This counter-intuitive finding has been explained by a considerable knowledge asymmetry which prevents customers from assessing the quality of the advice they receive. Absent a sufficient level of financial literacy, clients are forced to trust financial advice. Moreover, the lacking transparency about fee schedules and potential conflicts of interest created by sales-based incentives require a substantial leap of faith on the part of advisees when entrusting large sums of their money to advisors (e.g. Georgarakos & Inderst, 2011).

At the same time, however, owing to the integrity violations of many market players uncovered in the aftermath to the financial crisis, global trust in the banking industry has declined dramatically, making banks and financial services the least trusted industries by a long way (Guiso, 2010; Edelman, 2015). Taken together, individuals thus seem to trust their financial advisors whereas they have rather low levels of trust in the financial system in general.

In this study, we investigate how these two dimensions of trust formation, i.e. interpersonal trust in the advisor and broader trust in the financial industry, impact households' overall trust in financial advice. Given that the two trust components appear to have unique antecedents, analyzing their respective influence seems worthwhile in order to enhance our understanding about how trust

[1] See Lusardi and Mitchell (2014) for a recent review of the literature on financial literacy.

develops in the context of financial advice, i.e. a setting where clients have been found to be largely ignorant of conflicts of interest and thus are particularly vulnerable to opportunistic behavior exploiting their interpersonal trust in the financial advisor.[2]

Interestingly, however, while a number of studies in economics and marketing have examined contextual determinants of trust in advice that go beyond consumers' trust in the advisor (Moorman et al., 1993; Smith & Barclay, 1997; McMillan & Woodruff, 1999; Jeffries & Reed, 2000), prior research in household finance has focused on the role of interpersonal trust between advisee and advisor, thereby implicitly equating trust in financial advice with trust in the financial advisor. In fact, a review of the literature reveals that Grayson et al. (2008) are the only ones to explicitly allow for additional dimensions of trust formation in the financial services context and only recently, Monti et al. (2014) revisit the issue stating that "it would be interesting to see if other contextual cues would lead advisees to wisely choose a different mode of trust formation in an environment with less well aligned incentives to avoid the pitfalls of trusting senders misleading signals with harmful intent." (p. 1756). Yet, presumably owing to data limitations, they do not analyze this question empirically.

We fill this gap and extend the literature on the determinants of trust formation in the context of financial advice by applying the integrated conceptualization developed in Grayson et al. (2008) to take into account the potential impact of broad-scope trust for advisees' overall trust in financial advice. Using novel survey data obtained from the Panel on Household Finances (PHF) provided by the Deutsche Bundesbank, we contrast households' trust in financial advisors employed at community banks versus large banks, i.e. two bank types which have been shown to feature fundamentally different trust profiles (Hurley et al., 2014). This unique setting allows us to differentiate two layers of trust, i.e. narrow-scope trust towards a representative of a given financial services provider and broad-scope trust in the business context in which the financial services provider operates.

To preview our key results, we document that financial advice provided by large-bank advisors is significantly less likely to be trusted by the households surveyed in the PHF. Thus, our results prompt us to reject the notion that trust in financial advice is essentially equivalent to trusting one's financial advisor. Instead, we provide strong evidence in support of an integrated conceptualization of customers' trust in financial advice, which highlights the role of advisees' broad-

[2] See section 2.1 for a detailed discussion of the related evidence.

scope trust in the business context in which the provider of financial advice operates.

The remainder of this study is organized as follows. In section 2, we relate our work to prior research on trust in the context of financial advice and derive our hypotheses. Section 3 presents our data and descriptive statistics. In sections 4 and 5, we present and discuss our empirical results. Section 6 concludes.

2. Related research and hypothesis development

2.1. Interpersonal trust formation: The customer-advisor interaction

Theory on how trust is formed in an advice context posits that the alignment of advisor and consumer incentives is the most significant factor for consumer trust to develop at the interpersonal level (Yaniv & Kleinberger, 2000; Sniezek & Van Swol, 2001). However, despite a few contributions that either model clients as rational agents who are aware of the advisor's selling incentives (Calcagno & Monticone, 2015) or allow for them to vary in their understanding of the advisor's conflict of interest (Inderst & Ottaviani, 2012), empirical studies in the field overwhelmingly document that consumers do not possess the discernment to tell conflicted recommendations from unbiased advice. Based on a large-scale survey among six thousand investment advisees in eight EU member states, Chater et al. (2010), for instance, show that respondents are largely ignorant of conflicts of interest.

Instead, several studies document that advisees turn to salient factors when forming their impressions about the trustworthiness of the advisor. In an early study, Johnson and Grayson (2005) examine relationships between consumers and financial advisors and conceptualize trust as having cognitive and affective dimensions. While cognitive trust is knowledge-driven, affective trust arises from the confidence the client places in her advisor based on feelings generated by the level of care and concern which the advisor demonstrates. Given that a substantial knowledge asymmetry typically prevents customers from assessing the quality of the advice they receive, the authors highlight the role of affective trust in financial advice. This finding is corroborated in a comprehensive audit study by Mullainathan et al. (2013), in which trained mystery shoppers consult with financial advisors to discuss their portfolio composition. The authors report that, on average, advisors fail to debias the auditors and even encourage misconceptions which are in line with their own interests by reinforcing return chasing and promoting the reallocation of assets into actively managed funds with higher fees. Paradoxically enough, the majority of mystery shoppers nevertheless stated that they

would return to the advisors they consulted in order to obtain real-world recommendations even after they had learned about their self-interested catering strategies in the subsequent debriefing.

In a related study, Monti et al. (2014) survey retail investors at an Italian cooperative bank and show that actual investment decisions can be explained in large part by a simple heuristic based on how customers perceive the communication style of their financial advisors rather than by the features of the recommended investment products. Similarly, Agnew et al. (2014) document in an experimental setting that customers use advisors' professional credentials as a sign of expertise, but face severe difficulties discriminating fake credentials from real ones, which undoes the signal effect. Taken together, customers seem to be largely naïve to moral hazard issues when judging their advisors' trustworthiness, although recent research suggests that this mode of interpersonal trust formation - i.e. independent of fundamentals - may well be exploited by opportunistic advisors. Gennaioli et al. (2015) present a model in which trusted advisors do not correct investors' errors but instead have a strong incentive to cater to their biased beliefs. This prediction is supported by the experimental results in Agnew et al. (2014) who demonstrate that a customer's perception of her advisor's ability can be manipulated by using a simple strategy where confirming the client's preexisting view on an easy topic builds trust in the advisor which subsequently persists regardless of the quality of future advice.[3]

2.2. An integrated conceptualization of customer trust formation

Given that customers' trust in the financial advisor may not always be deserved and appears to be rather easily won by simple catering strategies, are there other levels of trust formation which determine peoples' overall trust in financial advice? While a number of studies in economics and marketing have examined contextual determinants of trust in advice that go beyond consumers' trust in the advisor (Moorman et al., 1993; Smith & Barclay, 1997; McMillan & Woodruff, 1999; Jeffries & Reed, 2000), Grayson et al. (2008) are the first to allow for additional dimensions of trust formation in the financial services context and conclude that trust in the advisor is not the same as trust in the advice. Instead, they find that customers are influenced not only by how much they trust a given company

[3] Note that Johnson and Grayson (2005) provide early anecdotal evidence in support of advisor catering: one of the financial advisors interviewed in the study states that "a tactic use by advisers to gain the trust of first-time customers is to recommend a product that saves the customer transaction fees and earns little or no commission for the adviser. The adviser informs the customer of this act of benevolence, which elicits an emotional bond of trust in the financial adviser." (p. 501).

and its representatives but also by how much they trust the broader context in which the market exchange is taking place. Accordingly, they present an integrated conceptualization of customer trust formation in the context of financial services, which distinguishes two layers of trust, i.e. narrow-scope trust at the interpersonal level and broad-scope trust in the business context in which a financial services provider operates. At this, interpersonal trust is narrow in scope because it only affects the relationship from which it has originated. Broad-scope trust, on the other hand, depends on the social context in which the relationship is maintained (Driscoll, 1978). While the research of Grayson et al. (2008) is somewhat related to ours, their focus lies on testing two rival sociological perspectives regarding the influence of customer trust in the broader context and they conclude that broad-scope trust and narrow-scope trust are complements rather than substitutes. By contrast, we are interested in how the two dimensions of trust, i.e. trust in the advisor and broader trust in the industry providing financial advisory services, impact advisees' overall trust in financial advice.

2.3. Trust profiles of community banks versus large banks

To capture the potential influence of broad-scope trust, we make use of a unique feature of our data, i.e. the fact that not only we know households' likelihood to trust the financial advice of their house bank, but also have information about the bank type to which it belongs. Given this setting, we are able to compare households' propensity to trust financial advice provided by advisors employed at community banks versus large banks, which, in a recent study by Hurley et al. (2014), have been shown to feature fundamentally different trust profiles. Hurley et al. (2014) apply the framework of customer trust developed by Grayson et al. (2008) to case study data and show that core elements of trustworthiness ingrained in the business model of community banks are missing in many large banks.

The first aspect addresses differences in the general value proposition of the two bank types to their respective customers as well as the wider society. Specifically, community banks have a mandate to serve the public interest (savings banks) and promote local economic development (cooperative banks), and earn significant "benevolence credits" through a number of community-building activities that are well-aligned with their business models. By contrast, large banks do not have a tradition of connecting their core business models to socially redeeming purposes but instead have predominantly been committed to maximizing shareholder value.

Second, regarding the sustainability of the business model, most community banks (as opposed to the majority of large banks) have accepted slower growth

in the run-up to the financial crisis so as to avoid venturing into lines of business where client conflicts were likely. Similarly, they refrained from securitizing their mortgage portfolios to show alignment with local borrowers.

Third and finally, the recent crises have uncovered substantial problems regarding the integrity and compliance of the business models of many large banks all over the world. Fraudulent behavior and deception such as 'robosigning' of mortgage contracts or the manipulation of interest rates revealed that in many cases, the original goal of many banks was to increase bonuses and the short-term market value, irrespective of the long term risk to stakeholders and the society. The fact that several large banks were eventually bailed out despite severe integrity violations not only undermined peoples' trust in the regulatory authorities but also further damaged the reputation of and trust in large banks.

2.4. Hypotheses

Hurley et al. (2014) conjecture that these differences in the respective business models have undermined peoples' broad-scope trust in large banks in the aftermath of the financial crisis of 2008 and thus provide us with an empirically testable implication. Combining the findings on interpersonal trust formation discussed in section 2.1 with the differences in the trust profiles of community banks as opposed to large banks, we investigate the respective roles of narrow-scope trust in the advisor and broad-scope trust in the business context as determinants of individuals' overall trust in financial advice. Specifically, we hypothesize that the considerable differences in strategy and culture of community banks versus large banks should manifest in significantly lower trust levels of clients advised at large banks in case broad-scope trust indeed plays a role in customer trust formation. If, however, trust in financial advice is essentially equivalent to trusting one's financial advisor, we should not observe material differences in trust levels of advisees at community banks and large banks, respectively.

3. Data

3.1. The Panel on Household Finances (PHF)

To obtain individuals' propensity to trust their advisors depending on what type of bank the latter are employed with, we draw on novel survey data on household finance and wealth in Germany provided by the Deutsche Bundesbank in the Panel on Household Finances (PHF) which is representative of the German population. Interviews with the 3,565 households sampled in the first wave of the PHF were conducted between September 2010 and July 2011 and questions cover

a wide range of items related to the household balance sheet including financial and non-financial assets as well as household debt. This information is then supplemented with demographic and psychological characteristics of the household members as well as a household-specific financial literacy score. Detailed variable descriptions are given in Appendix II-1.

Finally, the PHF features (a) survey weights to adjust for the oversampling of wealthy households during the data collection[4] and (b) multiple imputations in order to mitigate the issue of missing data due to item non-response. Following Bucher-Koenen and Ziegelmeyer (2014), we do not use imputed values for our dependent variables and thus omit the respective households from our final sample.[5]

For the subsample of households who have received financial advice within two years prior to being interviewed (N=965), we assess trust in the advice using the PHF items "Looking to the near future: How likely is it that your household will follow the advice provided by your house bank" with possible answers coded in a binary variable ("Rather likely." versus "Rather unlikely.")[6] and "To which banking group does your household belong?" ("Savings bank", "Cooperative bank", "Large bank", "Direct bank", "Other"). Straightforwardly, we classify savings banks and cooperative banks as community banks and contrast them with the group of large banks. Moreover, we exclude direct banks and other institutions since they do not offer retail financial advice. By explicitly relating the trust item to the respondent's primary relationship bank ('house bank'), the PHF captures the great majority of advised individuals in Germany: Hackethal et al. (2010) document that, unlike consumers in the US or the UK, German retail investors overwhelmingly report to seek financial advice at their house banks.

---

[4] We make use of the survey weights and the corresponding replicate weights to adjust point estimates as well as variance and standard error estimates in all our baseline analyses. In section 4.2.2, we analyze if this correction of the sampling design affects our main results.

[5] Note that for the independent variables, we use the average of the five imputed values provided in the data. For robustness, we re-estimate our main model using multiple imputations via Rubin's rule (Rubin, 1996). Results remain virtually unchanged and are reported in Appendix II-3.

[6] This approach of eliciting a client's trust in her financial advisor via the likelihood with which she heeds her recommendations seizes on the notion that trust should translate into behavioral manifestations of trust (Mayer et al., 1995) and follows Lachance and Tang (2012), who measure trust in financial advice based on the extent to which respondents to the National Financial Capability Survey (NFCS) agree to "accept what [the financial professional] recommends". Johnson and Grayson (2005) choose a similar trust construct by inquiring into the degree to which interviewees "have no reservations about acting on [their financial advisors'] advice".

Table II-1: Descriptive statistics

This table provides descriptive statistics for the households in our sample obtained from the Panel on Household Finances (PHF). The data are weighted and representative for Germany. The PHF is provided with multiple imputations which are estimated via Markov-Chain-Monte-Carlo method (Zhu & Eisele, 2013). We do not use multiple imputations for our dependent variable. For the remaining variables, we use the average of the five imputed data points. Appendix II-1 provides variable descriptions.

	Advised						Non-advised				
	N	Mean	SD	Min.	Median	Max.	N	Mean	SD	Diff.	t-test
Panel A: Financial variables											
TRUST_FA	929	0.615	0.487	0	1	1					
COMMUN_BANK	945	0.775	0.417	0	1	1	2,173	0.810	0.392	-0.035	1.37
FIN_LITERACY	955	2.550	0.686	0	3	3	2,243	2.447	0.757	0.103	1.53
FIN_WEALTH	965	67,444	168,875	0	31,800	5,000,000	2,287	23,730	77,142	43,714***	8.02
RISK_PROP	965	1.558	0.552	1	2	4	2,286	1.327	0.546	0.231***	7.22
Panel B: Sociodemographic characteristics											
GENDER	965	0.578	0.494	0	1	1	2,287	0.487	0.5	0.091***	3.2
MARRIED	965	0.535	0.499	0	1	1	2,287	0.491	0.5	0.044	1.58
AGE	965	51.9	16.8	19	50	90	2,287	52.6	18.0	-0.674	0.58
INCOME	965	2,727	1,806	350	2,300	40,000	2,287	2,178	2,453	549***	5.99
WEALTH	965	238,229	823,083	0	115,000	60,000,000	2,287	129,508	378,000	108,721***	4.98
EDU_HIGH	965	0.669	0.914	0	0	3	2,287	0.461	0.790	0.2071***	3.93
EMPL_SELF	965	0.054	0.227	0	0	1	2,287	0.070	0.256	-0.016	1.25
TRUST_GEN	964	5.565	1.912	0	5	10	2,283	5.342	2.185	0.22*	1.94

Table II-1 reports descriptive statistics of the variables which we include in our analysis. Specifically, the dummy variable TRUST_FA equals one for the 61.5% of respondents who considered it "rather likely" to implement the financial advice they had obtained, while the remaining 38.5% of households who had sought financial advice during the period under review stated that they were "rather unlikely" to follow it. Moreover, 77.5% of advisees report their house bank to be a community bank (in which case the dummy variable COMMUN_BANK takes the value one). Finally, Table II-1 shows that the demographic profile of advised respondents is only partly representative of the average household. Compared to the group of non-advised households, we find that they do not differ materially in terms of age, family and employment status as well as financial literacy levels and the likelihood of having a community bank as their house bank. At the same time, however, advised respondents on average dispose of substantially higher income and wealth, are more educated, more likely to be males, and have a greater risk appetite.

3.2. Demographic profiles of bank clienteles

Before we turn to explaining our key variable TRUST_FA, we explore our main sample of advised households in more detail and now use a multivariate

setting to investigate if the clienteles of community banks and large banks differ systematically with respect to their demographic profiles.

Table II-2: Demographic profiles of bank clienteles

This table reports average marginal effects of a series of probit regressions with COMMUN_BANK as the dependent variable. Appendix II-1 provides variable descriptions. Standard errors are reported below the coefficients in parentheses. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively.

	Dependent Variable: COMMUN_BANK		
	All	Advised	Non-advised
FIN_LIT	-0.0010	0.0212	-0.0050
	(0.0170)	(0.0343)	(0.0157)
FIN_WEALTH(log)	-0.0020	-0.0058	-0.0001
	(0.0043)	(0.0144)	(0.0048)
RISK_PROP	-0.0458**	-0.0658*	-0.0404*
	(0.0221)	(0.0353)	(0.0238)
GENDER	-0.0022	0.0173	-0.0022
	(0.0200)	(0.0433)	(0.0219)
MARRIED	-0.0013	0.0315	-0.0123
	(0.0264)	(0.0378)	(0.0280)
AGE_36-50	-0.0255	-0.0645	-0.0212
	(0.0377)	(0.0766)	(0.0383)
AGE_51-65	-0.0037	-0.0803	0.0231
	(0.0354)	(0.0742)	(0.0389)
AGE_65+	0.0345	-0.0593	0.0581
	(0.0379)	(0.0785)	(0.0402)
EDU_HIGH	-0.0606***	-0.0654***	-0.0569***
	(0.0111)	(0.0235)	(0.0138)
SELF_EMPL	0.0215	-0.1780*	0.0768***
	(0.0346)	(0.1014)	(0.0286)
WEALTH(log)	0.0029	0.0006	0.0038
	(0.0040)	(0.0126)	(0.0041)
INCOME(log)	-0.0081	0.0763**	-0.0372
	(0.0206)	(0.0369)	(0.0244)
N	3,067	935	2,129
Wald Chi^2	66.82	26.97	74.13
(p-value)	(0.000)	(0.008)	(0.000)

Table II-2 reports regressions of COMMUN_BANK on the various parameters capturing the financial situation of the households and the demographics available in the PHF, both for the full sample as well as for the subsamples of advised versus non-advised households. We observe a number of interesting results. First, both the basic demographic characteristics of advised households and their financials (with the exception of income) turn out insignificant in distinguishing between the clienteles of community banks and large banks. Second, the significant difference between community-bank clients and large-bank clients regarding their

general educational background does not translate into a relevant gap when it comes to their knowledge in financial matters as measured by the financial literacy score. Third and last, we note that the demographic profiles of non-advised versus advised households do not differ materially, suggesting that the comparability of the customer groups across the different bank pillars does not hinge upon whether or not they have sought advice in the past.

3.3. Trust determinants

Since we are the first to make use of the PHF survey for an analysis of households' trust in financial advice, we follow Lachance and Tang (2012) and start by developing a better understanding of our key variable TRUST_FA. To this end, we compare it to the generalized trust in people question (TRUST_GEN) which has been used in early studies relating trust and financial markets (e.g. Guiso et al., 2008; Georgarakos & Pasini, 2011) and is worded "Are you generally a person who trusts others or do you tend to be distrustful of others?" with possible scores ranging from 0 ("I do not trust others at all.") to 10 ("I trust others completely."). A direct comparison of the two interpersonal trust constructs allows us to learn more about consumers' trust in financial advisors by testing whether or not it is different from their trust in people in general. To facilitate comparison between the two items, we recode TRUST_GEN by means of a median split in order to have it on the same scale as TRUST_FA.[7] We estimate a probit model for each trust construct, regressing it on the available household characteristics.

Table II-3 reports the corresponding results. Interestingly, the goodness-of-fit statistics indicate that a model using a comprehensive set of households' basic demographics and financial parameters as inputs is not able to explain their general trust towards others as captured in TRUST_GEN (p-value of the corresponding Wald test equals 0.146).

[7] Note that OLS regression results using the unadjusted scale of TRUST_GEN (available upon request) produce qualitatively similar results.

Table II-3: Trust determinants

This table reports average marginal effects of probit regressions with TRUST_FA and TRUST_GENERAL as the dependent variables. Appendix II-1 provides variable descriptions. Standard errors are reported below the coefficients in parentheses. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively.

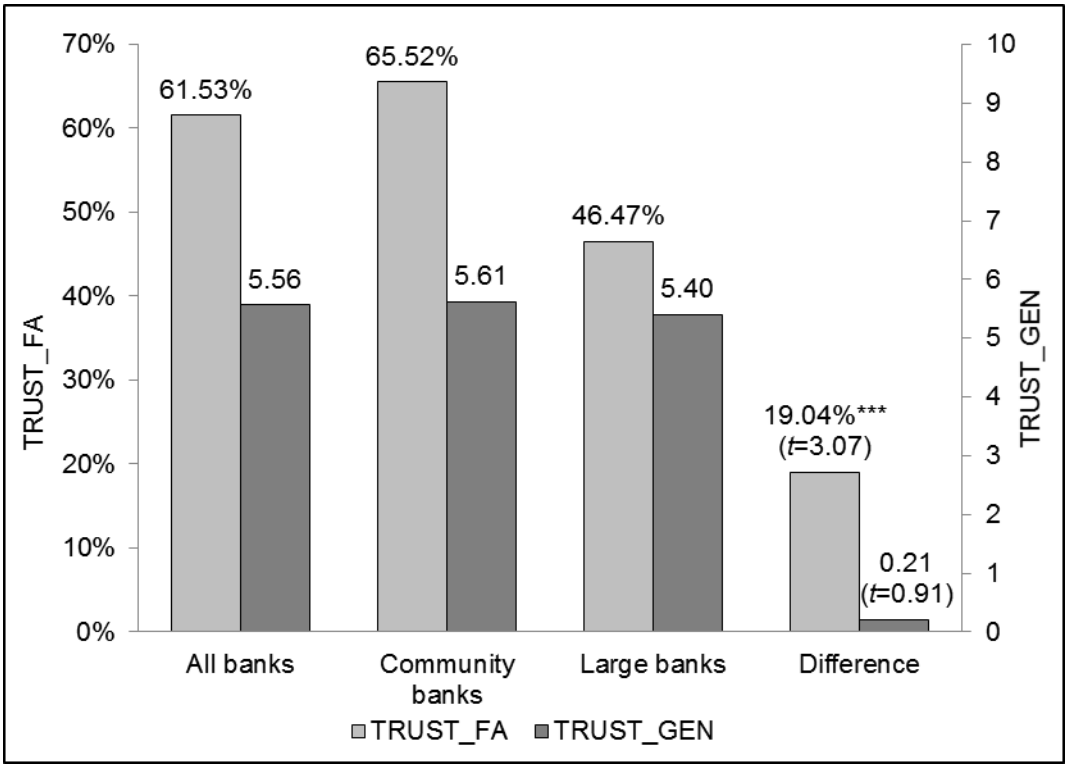
	TRUST_FA	TRUST_GEN
FIN_LIT	-0.0951**	0.0416
	(0.0448)	(0.0387)
FIN_WEALTH(log)	0.0431**	-0.0229
	(0.0169)	(0.0165)
RISK_PROP	0.0227	0.1321***
	(0.0403)	(0.0476)
GENDER	-0.0683	-0.0205
	(0.0462)	(0.0554)
MARRIED	0.1121**	0.0105
	(0.0530)	(0.0608)
AGE_36-50	-0.0476	-0.0506
	(0.0850)	(0.0878)
AGE_51-65	-0.1316	0.0483
	(0.0963)	(0.0973)
AGE_65+	-0.0323	0.0006
	(0.0843)	(0.0987)
EDU_HIGH	-0.0152	0.0256
	(0.0271)	(0.0260)
SELF_EMPL	-0.0855	-0.1816*
	(0.0905)	(0.0940)
WEALTH(log)	0.0164	-0.0079
	(0.0173)	(0.0151)
INCOME(log)	-0.0277	0.1127*
	(0.0575)	(0.0623)
N	919	918
Wald Chi^2	25.63	17.10
(p-value)	(0.012)	(0.146)

Turning to the determinants of TRUST_FA, we first observe a similar pattern for the demographic variables. Specifically, respondents' age, gender, education, self-employment as well as their income and aggregate wealth do not significantly impact the trust they have in their financial advisors. When looking at the parameters describing the households' financial situation, however, we find that households featuring above-average financial wealth and comparatively low financial literacy turn out to be significantly more likely to trust their financial advisors. Given that low financial literacy has been shown to further decrease advisees' ability to discern good from bad advice (Georgarakos & Inderst, 2011; Hackethal et al., 2012) and, at the same time, the damage from bad advice likely increases in financial wealth, this result highlights the importance of using a context-based

measure when analyzing the interpersonal trust component of customers' overall trust in financial advice.

Figure II-1: Trust in financial advice and general trust

This figure plots respondents' average levels of general trust towards other people (TRUST_GEN) as well as their trust towards financial advice they have received (TRUST_FA), thereby differentiating between the different bank clienteles. *** indicates statistical significance at the 1% level.



To capture the impact of broad-scope trust, we contrast the trust levels of customers at community banks and large banks, respectively. Figure II-1 plots the corresponding results and shows that, while general trust levels are virtually identical across the different bank pillars, trust in financial advisors differs sharply between clients at community banks versus large banks. Specifically, community-bank advisees are as much as 19 percentage points more likely to follow the advice of the financial professionals they have consulted (65.5% versus 46.5%, t=3.07). Thus, our initial univariate comparison supports the hypothesis that the fundamental differences in the trustworthiness of the business models of community banks versus large banks manifest in significantly lower trust levels of clients advised at large banks. This implies that broad-scope trust is important for cus-

tomer trust formation and rejects the idea that trust in financial advice is essentially equivalent to trusting one's financial advisor. In what follows, we investigate if this difference persists in a multivariate setting.

4. Results

4.1. Main results

To examine the impact of the respondents' affiliation to either of the two bank types on their likelihood to trust financial advice, we estimate simple probit models whose results we report in Table II-4.

In what follows, we briefly discuss our findings in light of prior evidence on particularly robust determinants of financial advice other than trust, i.e. financial literacy and financial wealth as well as age. First, our finding that financial literacy is negatively related with trust in financial advice ties in with robust evidence presented in a number of studies including Lachance and Tang (2012) and Calcagno and Monticone (2015) and, for the German market, Hackethal et al. (2010), Bucher-Koenen and Koenen (2015), and Stolper (2016), who all document that individuals are less likely to implement the advice given to them when their financial sophistication is higher. To rationalize the adverse impact of financial knowledge on trust in financial advice, it is argued in the literature that increased financial sophistication involves the competence to question the advice along with better skills to process information relevant for decision-making privately. Thus, households seem to become more critical as to the value proposition offered by financial advisors once they have gathered a sufficient degree of financial literacy. Again, this finding is particularly relevant when turning to the less financially knowledgeable customers who do not possess an outside option and need to rely on the recommendations they receive from their advisors. Clearly, these clients are particularly vulnerable to opportunistic behavior exploiting their trust in the financial advisor. Moreover, the positive impact of financial wealth on households' propensity to trust their advisors supports the findings in Bhattacharya et al. (2012) and Lachance and Tang (2012), who show that individuals who are wealthier in financial assets are more likely to follow the recommendations of their advisors. On the one hand, this may be justified in light of survey evidence of Tilmes and Jakob (2012), who document that the discretionary power of advisors typically increases in the financial assets they are entrusted with by a given advisee. Combined with their finding that advisors' self-reported perception of conflicts between their own interests and the customer benefit on average decreases in the individual discretion they are conceded when advising their clients, the observed

increase in trust among advisees with greater financial wealth might as well be "earned".

Table II-4: Bank clienteles and trust in financial advice

This table reports average marginal effects of a series of probit regressions featuring TRUST_FA as the dependent variable. Column (1) reports univariate results for our key explanatory variable COMMUN_BANK. Column (2) reports the results of a multivariate regression including all control variables. For ease of comparison, column (3) replicates column (1) of Table II-3, i.e. a multivariate regression on the controls only. Appendix II-1 provides variable descriptions. Standard errors are reported below the coefficients in parentheses. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively.

	Dependent Variable: TRUST_FA			
COMMUN_BANK	0.1904***	0.1936***		
	(0.0670)	(0.0658)		
TRUST_GEN		0.0329***	0.0349***	
		(0.0125)	(0.0128)	
FIN_LIT		-0.1072***	-0.1019**	-0.0951**
		(0.0393)	(0.0423)	(0.0448)
FIN_WEALTH(log)		0.0451***	0.0455***	0.0431**
		(0.0168)	(0.0172)	(0.0169)
RISK_PROP		0.0128	0.0027	0.0227
		(0.0434)	(0.0432)	(0.0403)
GENDER		-0.0540	-0.0648	-0.0683
		(0.0470)	(0.0450)	(0.0462)
MARRIED		0.1124**	0.1085**	0.1121**
		(0.0526)	(0.0522)	(0.0530)
AGE_36-50		-0.0284	-0.0325	-0.0476
		(0.0832)	(0.0844)	(0.0850)
AGE_51-65		-0.1082	-0.1262	-0.1316
		(0.0929)	(0.0928)	(0.0963)
AGE_65+		-0.0237	-0.0323	-0.0323
		(0.0828)	(0.0830)	(0.0843)
EDU_HIGH		-0.0073	-0.0176	-0.0152
		(0.0256)	(0.0264)	(0.0271)
SELF_EMPL		-0.0416	-0.0637	-0.0855
		(0.0885)	(0.0890)	(0.0905)
WEALTH(log)		0.0166	0.0175	0.0164
		(0.0162)	(0.0171)	(0.0173)
INCOME(log)		-0.0432	-0.0402	-0.0277
		(0.0541)	(0.0558)	(0.0575)
N	909	898	918	919
Wald Chi^2	8.03	61.61	38.66	25.63
(p-value)	(0.005)	(0.000)	(0.000)	(0.012)

On the other hand, however, a less favorable interpretation of the results could be that advisors put more effort in catering strategies to build client trust since generating credibility is arguably more profitable in case of financially wealthier customers. Under this scenario, an increase in trust levels does not necessarily

reflect better advice and potential disadvantages from receiving self-interested recommendations would become worse the higher the stakes of the advisee.

Third, neither age nor risk propensity feature explanatory power in our sample. This is somewhat at odds with the results in Mullainathan et al. (2013) and Lachance and Tang (2012) who report that elder advisees are less trustful of their advisors. Similar to the interpretation of the adverse effect of financial literacy on trust, these studies propose that elder clients are more experienced in financial matters and thus also more skeptical regarding the benefits of financial advice. We cannot confirm this relation for our sample of advised households.

Finally, we note that the coefficients of the previously identified drivers of trust in financial advice reported in Table II-3 and, for ease of comparison, replicated in the rightmost column of Table II-4, are virtually unchanged once we add customers' bank type as an additional trust determinant. Thus, our key variable TRUST_FA introduces a new dimension of the client-advisor trust formation process which has not yet been captured by prior explanations. Taken together, the results presented in this section prompt us to reject the notion that that trust in financial advice is essentially equivalent to trusting one's financial advisor. Instead, we provide strong evidence in support of an integrated conceptualization of customers' trust in financial advice, which highlights the role of advisees' broad-scope trust in the business context in which the provider of financial advice operates.

4.2. Robustness analysis

4.2.1. Potential endogeneity of bank choice

To examine the robustness of our key findings, we consider potential endogeneity concerns when studying households' choice of their house bank (community bank versus large bank) as well as their likelihood to trust the financial advice they receive at their house bank. Since respondents in our sample are asked to express their propensity to follow the financial advice at their house bank conditional on having received advice at that bank, reverse causality (i.e. bank choice endogenously determined by a given household's trust in their financial advisor) is rather unlikely to be an issue in our analysis. However, we consider the possibility that an unobserved variable simultaneously drives both the selection of the type of house bank and the propensity to trust the financial advisor employed at the chosen type of bank. Given the evidence in Mullainathan et al. (2008), who emphasize that banks tend to advertise their trustworthiness rather than their performance, one such omitted factor might be a bank's reputation. If consumers are trustful towards the bank as an organization and choose to become a customer

of the bank as a result thereof, chances are that the bank's reputation positively affects their perception of the trustworthiness of the advisor working for that bank, too.[8]

We address this potential source of endogeneity by means of an two-stage instrumental variables (IV) regression approach featuring the two instruments RURAL and JOINT_DECISION, where RURAL represents a dummy variable that equals one if the respondent lives in a small municipality as opposed to a city. Similarly, JOINT_DECISION takes a value of one if household members report to decide financial matters jointly, and zero otherwise.

We argue that, since branches of large banks are much less densely distributed in the rural regions of Germany than are branches of community banks, households living in rural areas are likely to be limited in their choice options when selecting their house bank. Likewise, joint decision making presumably requires an increased organizational effort on the part of the advised household, e.g. due to the fact that all decision makers wish to attend the personal meetings with the advisor. Hence, households who live in rural areas and whose members jointly care about their household finances are arguably more likely to choose a house bank close to their place of residence. Given the much higher branch density of community banks outside of Germany's larger cities, these households should thus be more likely to be advised at a savings or co-operative bank as compared to a large bank.

Consequently, our two instrumental variables should both be highly correlated with the potentially endogenous variable COMMUN_BANK. Similarly, neither living in a rural area nor joint financial decision-making should have an impact on a household's propensity to trust their financial advisor, such that both instruments can reasonably be assumed to be uncorrelated with the error term of the first-stage regression.

[8] Ideally, we would of course want to control for bank reputation when analyzing the impact of clients' trust in financial advice. Owing to the aggregation level of the PHF data, however, we do not have any information at the level of the individual bank.

Table II-5: Robustness – Potential endogeneity of bank choice

This table reports the results of an IV regression along with the corresponding test statistics. The upper part of Table II-5 reports the first-stage estimates of a linear probability model estimated via GMM. The lower part of Table II-5 presents the second-stage estimates of the linear probability model. Robust standard errors are reported in brackets. Additional (control) variables are used but not reported. Appendix II-1 provides variable descriptions. Standard errors are reported below the coefficients in parentheses. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively.

Panel A: First stage (Dependent variable: COMMUN_BANK)		
RURAL	0.2410***	
	(0.0485)	
JOINT_DEC	0.0927**	
	(0.0425)	
	(other regressors suppressed)	
N	898	
R^2	0.113	
F-test excl. instr. (p-value)	15.11	(0.000)
Hansen J statistic (p-value)	0.765	(0.382)
Endogeneity test (p-value)	2.513	(0.113)
Panel B: Main equation (Dependent variable: TRUST_FA)		
COMMUN_BANK	0.5187**	
	(0.2159)	
	(other regressors suppressed)	
N	898	
R^2	0.046	

Panel A of Table II-5 reports the results of a re-estimation of our main model allowing for potential endogeneity as specified above. The estimates obtained from a first-stage regression indicate that both instruments are strongly correlated with COMMUN_BANK. Additional test statistics provided in the lower part of Table II-5 show that the IV model does not suffer from a weak instruments problem (F-test of excluded instruments: p=0.000) and provide evidence supporting the instruments' validity (Hansen J statistic: p=0.382). Finally, a formal endogeneity test supports the null hypothesis that households' bank choice is exogenous at all conventional significance levels (p=0.113). In summary, we thus conclude that the relationship between households' bank choice and their propensity to trust their financial advisor is robust to potential endogeneity.

4.2.2. Correction of standard error estimates

Recall that wealthy households are oversampled in the PHF data. To control for the oversampling, i.e. to provide adjusted point estimates which are representative of the German population, the data features survey weights which counterbalance the unequal selection probabilities caused by the biased sampling design. As recommended when using the PHF, we take these survey weights to adjust point estimates as well the corresponding replicate weights to adjust variance and standard error estimates in our main analysis. In order to assess if the weighting method applied to the standard error estimates potentially affects our results, we check for robustness by re-estimating our baseline model using (a) a Taylor linearization as an alternative weighting technique and (b) (unweighted) robust standard errors. Appendix II-2 documents the corresponding results and shows that they turn out virtually identical regardless of which correction method is applied. Specifically, the standard error of our key explanatory variable COMMUN_BANK is largest (albeit not materially different in magnitude) for the recommended correction method featuring replicate weights. This indicates that, if anything, our baseline model slightly understates the statistical significance of the effect of households' bank choice on their likelihood to trust their financial advisor, and we conclude that our main results prove robust to alternative methods of correcting the standard error estimates, too.

5. Discussion

5.1.1. Implications

Our results highlight the importance of establishing a climate of trust in the generic business context so as to enhance customers' propensity to trust financial advice which is nonetheless provided by individual representatives of specific players in that industry. Because of this collective goods problem, managers at large banks may refrain from investing in the development of broad-scope trust, since all banks within that group would benefit from higher levels of broad-scope trust. However, the substantially higher trustworthiness of German community banks among advisees suggests that the support of umbrella organizations (Deutscher Sparkassen- und Giroverband (DSGV) in case of savings banks and Bundesverband der Deutschen Volksbanken und Raiffeisenbanken (BVR) for cooperative banks, respectively) can be a worthwhile investment in the development of broad-scope trust. Moreover, Grayson et al. (2008) show that firm trust is essential even in a trusted environment, implying that banks must still provide the

means to establish narrow-scope trust before they can fully benefit from the customer attitudes and behaviors that are fostered by trust. Thus, for any given bank, the possibility to free ride on their competitors' investments in broad-scope trust is limited.

Similarly, effective regulation which enforces industry standards and codes of conduct may be a fruitful avenue to foster advisees' trust in the broader business context of large banks. Still, many managers oppose industry regulation for fear of precluding their firms from profitable business activities. Clearly, however, while not supporting government authorities and umbrella associations may eventually result in reduced regulatory requirements, this strategy is also unlikely to improve clients' broad-scope trust and hence negatively feeds back into their trust in financial advice.

5.1.2. Limitations and directions for future research

While the survey data provided in the Panel on Household Finances (PHF) allows us to draw our conclusions from a representative sample of clients across the entire universe of community banks and large banks throughout Germany, our empirical analysis has two potential shortcomings worth mentioning. First, we capture broad-scope trust by contrasting customer trust in financial advisors employed at community banks versus large banks. While the two bank types have been shown to feature fundamentally different trust profiles (e.g. Hurley et al. 2014), we do not claim our methodological approach to be ideal. Even though dummy variables have widely been used to proxy for broad-scope trust (e.g. McMillan & Woodruff, 1999; Guseva & Rona-Tas, 2001), a categorical differentiation presents a rather coarse measure of clients' trust in the broader context in which the advisor operates. Since customers within a given bank group likely vary in their levels of broad-scope trust, the respective group mean may be an inaccurate representation of the broad-scope trust of a given individual in that group. Unfortunately, we lack the interval data to address this drawback. However, the difference in trust levels between the two bank groups is so large in magnitude that we are confident that our results prove economically meaningful for the majority of households under review. Still, eliciting more nuanced perceptions of individuals' trust in the broader business context in which different bank groups operate presents a worthwhile avenue for further research and should improve the quality of future metrics of broad-scope trust.

A second potential limitation of this study is that idiosyncracies pertaining to our sample drive the observed effects. Specifically, the period under review which we examine in this study was marked by a dramatic loss of trust in large banks in the aftermath of the global financial crisis (Guiso, 2010; Hurley et al., 2014).

Thus, our findings might not be generalizable to other, more regular market cycles. However, the most recent wave of the Chicago Booth/Kellogg School Financial Trust Index which elicits the percentage of people trusting various types of banks suggests otherwise (Sapienza & Zingales, 2016). The survey explicitly differentiates between peoples' trust in credit unions and local banks (i.e. community banks) as opposed to their trust in national banks (i.e. large banks) and finds that while in December 2015, 59% and 61% of respondents report to trust credit unions and local banks, respectively, this share amounts to only 32% for the group of national banks. Thus, even though trust in both bank groups has slightly increased ever since 2009, the substantial gap in trustworthiness persists and has narrowed only marginally to about 28 percentage points since it peaked at roughly 37 percentage points in 2011. Based on this recent evidence corroborating the trust gap between community banks and large banks along with the fact that most large banks have largely maintained their business models ever since the financial crisis (Tilmes & Jakob, 2012), our key results should be robust to the sample period under review. However, overcoming this data limitation also makes a good candidate for future research on trust formation in the context of financial advice.

6. Conclusion

Without the confidence and financial literacy to bank autonomously, most households must trust financial advisors to gain access to the ever complex market for financial products and services. Consequently, learning about how clients form impressions about the trustworthiness of their advisors is key to understanding the customer-advisor relationship and, given the vulnerability of most households to opportunistic behavior of their advisors, addresses a matter of great relevance.

By contrasting clients' trust in the services of financial advisors employed at two banks types with fundamentally different trust profiles, i.e. community banks and large banks, this study contributes to the literature on customer trust formation in financial advisory services and provides evidence supporting the notion that advisees take into consideration the broader context when assessing the trustworthiness of the financial advice they receive. Thus, our results prompt us to reject the notion that trust in financial advice is essentially equivalent to trusting one's financial advisor. Instead, we provide strong evidence in support of an integrated conceptualization of customers' trust in financial advice, which shows that trust formation is influenced not only by the actions of an individual organization and its representatives but also by the broader trust profile of the business

model the organization commits to. Managerial implications include the potentially positive effect of an investment in professional associations and the support and enforcement of regulatory standards in order to enhance clients' broad-scope trust in financial advice. We hope that this study stimulates further research on the antecedents of trust formation in the context of financial advice, ideally with a focus on determinants outside the client-advisor interaction.

7. References

Agnew, J., Bateman, H., Eckert, C., Iskhakov, F., Louviere, J. J., & Thorp, S. (2014). Individual judgment and trust formation: An experimental investigation of online financial advice. *Working Paper*.

Bergstresser, D., Chalmers, J. M. R., & Tufano, P., (2009). Assessing the costs and benefits of brokers in the mutual fund industry. *Review of Financial Studies*, 22(10), 4129-4156.

Bhattacharya, U., Hackethal, A., Kaelser, S., Loos, B., & Meyer, S. (2012). Is unbiased financial advice to retail investors sufficient? Answers from a large field study. *Review of Financial Studies*, 25(4), 975-1032.

Bucher-Koenen, T. & Koenen, J. (2015). Do seemingly smarter consumers get better advice? *Working Paper*.

Bucher-Koenen, T., & Ziegelmeyer, M. (2014). Once burned, twice shy? Financial literacy and wealth losses during the financial crisis. *Review of Finance*, 18(6), 2215-2246.

Calcagno, R., & Monticone, C. (2015). Financial literacy and the demand for financial advice. *Journal of Banking & Finance*, 50, 363-380.

Chater, N., Huck, S., & Inderst, R. (2010). Consumer decision-making in retail investment services : A behavioural economics perspective. Report to the European Commission/SANCO (2010).

Driscoll, J.W. (1978). Trust and Participation in Organizational Decision Making as Predictors of Satisfaction. *Academy of Management Journal*, 21(1), 44-56.

DSGV (2014). DSGV Vermögensreport 2014. *Report*. Available at: www.dsgv.de/de/fakten-und-positionen/publikationen/vermoegensbarometer.html.

Edelman (2015). The 2015 Edelman Trust Barometer. Available at: http://www.edelman.com/insights/intellectual-property/2015-edelman-trust-barometer/.

Gennaioli, N., Shleifer, A., & Vishny, R. (2015). Money doctors. *The Journal of Finance*, 70(1), 91-114.

Georgarakos, D., & Inderst, R. (2011). Financial advice and stock market participation. *Working Paper*.

Georgarakos, D., & Pasini, G. (2011). Trust, sociability, and stock market participation. *Review of Finance*, 15(4), 693-725.

Grayson, K., Johnson, D., & Chen, D.-F.R. (2008). Is Firm trust essential in a trusted environment? How trust in the business context influences customers. *Journal of Marketing Research*, 45(2), 241-256.

Guiso, L. (2010). A Trust-driven financial crisis - Implications for the future of financial markets. *EEAG Report on the European Economy*, 53-70.

Guiso, L., Sapienza, P., & Zingales, L. (2008). Trusting the Stock Market. *The Journal of Finance*, 63(6), 2557-2600.

Guseva, A., & Rona-Tas, A. (2001). Uncertainty, risk, and trust: Russian and American credit card markets compared. *American Sociological Review*, 66(5), 623-646.

Hackethal, A., Haliassos, M., & Jappelli, T. (2012). Financial advisors: A case of babysitters? *Journal of Banking and Finance*, 36(2), 509-524.

Hackethal, A., Inderst, R., & Meyer, S., (2010). Trading on Advice. *Working Paper*.

Hurley, R., Gong, X., & Waqar, A. (2014). Understanding the loss of trust in large banks. *International Journal of Bank Marketing*, 32, 1-33.

Inderst, R., & Ottaviani, M. (2012). How (not) to pay for advice: A framework for consumer financial protection. *Journal of Financial Economics*, 105(2), 393-411.

Jeffries, F.L., & Reed, R. (2000). Trust and adaptation in relational contracting. *Academy of Management Review*, 25(4), 873-882.

Johnson, D., & Grayson, K. (2005). Cognitive and affective trust in service relationships. *Journal of Business Research*, 58(4), 500-507.

Lachance, M.-E., & Tang, N. (2012). Financial advice and trust. *Financial Services Review*, 21(3), 209-226.

Lusardi, A., & Mitchell, O. S. (2014). The economic importancce of financial literacy: Theory and evidence. *Journal of Economic Literature*, 52(1), 5-44.

Mayer, R.C., Davis, J.H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709-734.

McMillan, J., & Woodruff, C. (1999). Interfirm relationships and informal credit in Vietnam. *The Quarterly Journal of Economics*, 114(4), 1285-1320.

Monti, M., Pelligra, V., Martignon, L., & Berg, N. (2014). Retail investors and financial advisors: New evidence on trust and advice taking heuristics. *Journal of Business Research*, 67(8), 1749-1757.

Moorman, C., Deshpandé, R., & Zaltman, G. (1993). Factors affecting trust in market research relationships. *Journal of Marketing*, 57(1), 81-101.

Mullainathan, S., Noeth, M., & Schoar, A. (2013). The market for retirement financial advice. *Working Paper*.

Mullainathan, S., Schwartzstein, J., & Shleifer, A. (2008). Coarse thinking and persuasion. *The Quarterly Journal of Economics*, 123(2), 577-619.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473-489.

Sapienza, P., & Zingales, L. (2016). Wave 24 of the Chicago Booth/Kellogg School Financial Trust Index reveals heightened public trust in local banks, credit unions. Available at: www.financialtrustindex.com.

Smith, J. B., & Barclay, D. W. (1997). The effects of organizational differences and trust on the effectiveness of selling partner relationships. *Journal of Marketing*, 61(1), 3.

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes*, 84(2), 288-307.

Stolper, O. (2016). It takes two to tango: Households' response to financial advice and the role of financial literacy. *Working Paper*.

Tilmes, P. R., & Jakob, R. (2012). Anlageberatung aus Sicht der Berater - eine Herausforderung zwischen Kunde, Kreditinstitut und Finanzmarktaufsicht.

Von Gaudecker, H. M. (2015). How does household portfolio diversification vary with financial literacy and financial advice? *The Journal of Finance*, 70(2), 489-507.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83(2), 260-281.

8. Appendix

Appendix II-1: Variable descriptions

This table describes the variables used in the study in alphabetical order.

Name	Description
AGE_36-50	Dummy variable that equals one if the respondent is aged 36 to 50 years. Zero otherwise.
AGE_51-65	Dummy variable that equals one if the respondent is aged 51 to 65 years. Zero otherwise.
AGE_65+	Dummy variable that equals one if the respondent is aged more than 65 years. Zero otherwise.
COMMUN_BANK	Dummy variable that equals one if the respondent has received financial advice at a community bank (i.e. savings bank or cooperative bank) and zero otherwise. Corresponding PHF item: "To which banking group does your household's house bank belong?" 1–Savings bank/Landesbank; 2–Cooperative bank; 3–Commercial bank
EDU_HIGH	Ordinal variable that describes the respondent's highest degree of education: 1–Higher education entrance qualification; 2–University degree; 3–Ph.D. or higher qualification. Zero otherwise.
FIN_LITERACY	Ordinal variable that measures the number of correctly answered financial literacy questions. Corresponding PHF items: *Question 1: Compound interest effect* "Let us assume that you have a balance of 100 EUR on your savings account. This balance bears interest at a rate of 2% per year and you leave it for 5 years on this account. How high do you think your balance will be after 5 years?" 1–More than 102 EUR [correct]; 2–Exactly 102 EUR; 3–Less than 102 EUR *Question 2: Inflation* "Let us assume that your savings account bears interest at a rate of 1% per year and the rate of inflation is 2% per year. Do you think that in one year's time the balance on your savings account will buy the same as, more than, or less than today?" 1–More than today; 2–The same as today; 3–Less than today [correct] *Question 3: Diversification* "Do you agree with the following statement: 'Investing in shares of a company is less risky than investing in a fund containing shares of similar companies'?" 1–Agree; 2–Disagree [correct]
FIN_WEALTH	Continuous variable that measures the households' financial wealth (EUR).
GENDER	Dummy variable that equals one if the respondent is male, zero for female.
INCOME	Continuous variable that measures the household's monthly income (EUR).
JOINT_DEC	Dummy variable that equals one if the household members decide financial matters jointly. Zero otherwise. Corresponding PHF item: "In general, how does your household make investment decisions?" 1–Generally each person in the household makes their own decisions; 2–We decide important things together; 3–One household member decides for the whole household; 4–Depends
MARRIED	Dummy variable that equals one if the respondent is married, zero otherwise.
RISK_PROP	Ordinal variable that measures the respondents' propensity to take financial risks. Corresponding PHF item: "Which of the following statements comes closest to describing the attitude to risk when your household makes savings or investment decisions?" 1–We are not willing to take any financial risks; 2–We take average financial risks expecting to earn average returns; 3–We take above-average financial risks expecting to earn above-average returns; 4–We take substantial financial risks expecting to earn substantial returns
RURAL	Dummy variable that equals zero (one) if the respondent lives in a city (small municipality).
SELF_EMPL	Dummy variable that equals one if the respondent is self-employed or entrepreneur. Zero otherwise.
TRUST_GEN	Ordinal variable that measures the respondents' trust on a scale from 0 to 10. Corresponding PHF-item: "Are you generally a person who trusts others or do you tend to be distrustful of others?" 0–"I do not trust others at all."; [...]; 10–"I trust others completely."
TRUST_FA	Dummy variable that equals one if the respondent reports to be likely to trust the financial advice provided by his house bank in the future (conditional on having used financial advice in the past two years prior to being interviewed). Corresponding PHF item: "Looking to the near future: How likely is it that your household will follow the advice provided by its house bank?" 1–"Rather likely."; 2–"Rather unlikely."
WEALTH	Continuous variable that measures the household's gross wealth (EUR).

Appendix II-2: Robustness - Correction of standard error estimates

This table reports average marginal effects of probit regressions with COMMUN_BANK as the dependent variable. Appendix II-1 provides variable descriptions. Standard errors are reported below the coefficients in parentheses. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively.

	Dependent Variable: TRUST_FA		
	Replicate weights	Taylor Linearization	No adjustment
COMMUN_BANK	0.1936***	0.1936***	0.1936***
	(0.0658)	(0.0588)	(0.0558)
TRUST_GEN	0.0329***	0.0329***	0.0329**
	(0.0125)	(0.0124)	(0.0129)
FIN_LIT	-0.1072***	-0.1072***	-0.1072***
	(0.0393)	(0.0323)	(0.0344)
FIN_WEALTH(log)	0.0451***	0.0451***	0.0451***
	(0.0168)	(0.0152)	(0.0149)
RISK_PROP	0.0128	0.0128	0.0128
	(0.0434)	(0.0376)	(0.0422)
GENDER	-0.0540	-0.0540	-0.0540
	(0.0470)	(0.0375)	(0.0461)
MARRIED	0.1124**	0.1124**	0.1124**
	(0.0526)	(0.0518)	(0.0517)
AGE_36-50	-0.0284	-0.0284	-0.0284
	(0.0832)	(0.0757)	(0.0762)
AGE_51-65	-0.1082	-0.1082	-0.1082
	(0.0929)	(0.0834)	(0.0828)
AGE_65+	-0.0237	-0.0237	-0.0237
	(0.0828)	(0.0759)	(0.0783)
EDU_HIGH	-0.0073	-0.0073	-0.0073
	(0.0256)	(0.0243)	(0.0256)
SELF_EMPL	-0.0416	-0.0416	-0.0416
	(0.0885)	(0.0767)	(0.0804)
WEALTH(log)	0.0166	0.0166	0.0166
	(0.0162)	(0.0140)	(0.0135)
INCOME(log)	-0.0432	-0.0432	-0.0432
	(0.0541)	(0.0502)	(0.0486)
N	898	898	899
Wald Chi^2	61.61	55.74	57.95
(p-value)	(0.000)	(0.000)	(0.000)

Appendix II-3: Robustness - Multiple imputations via Rubin's rule

This table reports average marginal effects of a series of probit regressions which replicate our main results using multiple imputations via Rubin's rule (Rubin, 1996). Appendix II-1 provides variable descriptions. Standard errors are reported below the coefficients in parentheses. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively.

	Dependent Variable: TRUST_FA			
COMMUN_BANK	0.1904***	0.1935***		
	(0.0670)	(0.0658)		
TRUST_GEN		0.0328***	0.0348***	
		(0.0125)	(0.0128)	
FIN_LIT		-0.1070***	-0.1017**	-0.0949**
		(0.0393)	(0.0424)	(0.0448)
FIN_WEALTH(log)		0.0442**	0.0446**	0.0423**
		(0.0177)	(0.0179)	(0.0176)
RISK_PROP		0.0132	0.0032	0.0230
		(0.0437)	(0.0436)	(0.0406)
GENDER		-0.0538	-0.0646	-0.0681
		(0.0473)	(0.0452)	(0.0464)
MARRIED		0.1120**	0.1080**	0.1116**
		(0.0527)	(0.0523)	(0.0531)
AGE_36-50		-0.0283	-0.0324	-0.0474
		(0.0832)	(0.0843)	(0.0849)
AGE_51-65		-0.1069	-0.1249	-0.1303
		(0.0928)	(0.0927)	(0.0961)
AGE_65+		-0.0222	-0.0307	-0.0309
		(0.0828)	(0.0831)	(0.0843)
EDU_HIGH		-0.0072	-0.0175	-0.0152
		(0.0257)	(0.0264)	(0.0272)
SELF_EMPL		-0.0409	-0.0630	-0.0848
		(0.0886)	(0.0891)	(0.0906)
WEALTH(log)		0.0164	0.0173	0.0162
		(0.0162)	(0.0169)	(0.0171)
INCOME(log)		-0.0414	-0.0384	-0.0261
		(0.0540)	(0.0558)	(0.0574)
N	909	898	918	919
Wald Chi^2	8.03	60.02	37.93	25.10
(p-value)	(0.005)	(0.000)	(0.000)	(0.014)

III. When do households fail to repay their debt? The role of gender and financial literacy

Co-authors: Tobias Meyll, Andreas Walter
Own share: 70%

When do households fail to repay their debt? The role of gender and financial literacy

Tobias Meyll[a] Thomas Pauls[b] Andreas Walter[c]

Abstract - We study the role of gender and financial literacy for household over-indebtedness. Our results indicate that financially illiterate women restrain themselves from the debt markets. Those women who hold debt are significantly better in coping with their debt burdens compared to men, as they are dramatically less often over-indebted, particularly when it comes to unsecured consumer debt. Further, for both genders, we find that financial literacy significantly reduces over-indebtedness and show this effect to be robust against potential endogeneity.

Keywords: Household finance, Over-indebtedness, Household debt behavior, Financial literacy, Gender

JEL-Codes: D03, D12, E21

[a] Department of Financial Services, University of Gießen, Licher Str. 74, 35394 Gießen, Germany. Tobias.Meyll@wirtschaft.uni-giessen.de.

[b] Department of Financial Services, University of Gießen, Licher Str. 74, 35394 Gießen, Germany. Thomas.Pauls@wirtschaft.uni-giessen.de.

[c] Department of Financial Services, University of Gießen, Licher Str. 74, 35394 Gießen, Germany. Andreas.Walter@wirtschaft.uni-giessen.de.

1. Introduction

Whenever households have to conduct financial decisions, a profound understanding of basic financial concepts, commonly referred to as financial literacy, is of vital importance. Financially illiterate households are repeatedly found to conduct less favorable financial decisions. For example, they less frequently plan for their retirement (Bucher-Koenen & Lusardi, 2011; Lusardi & Mitchell, 2008), have lower capital market participation rates (van Rooij et al., 2011), hold less diversified portfolios (von Gaudecker, 2015), and in general, are more prone to miscellaneous investment mistakes (Lusardi & Mitchell, 2014; Stolper & Walter, 2017). Unfortunately, financial illiteracy is found to be quite widespread throughout the population and women seem to be particularly affected. Lusardi and Mitchell (2008) find that women are less financially literate and less likely to plan their financials compared to men, and thus, less prepared for their retirement. In a subsequent study, van Rooij et al. (2011) confirm women's lower financial literacy and find women to participate less often in financial markets. Almenberg and Dreber (2015) confirm their results and find that a lack of financial literacy explains a significant part of the lower stock market participation of women.

Recently, literature has begun to elaborate on the role of financial literacy and household debt behavior, whereby the debtors' gender received only sparse attention yet. Lusardi and Scheresberg (2013) highlight that great shares of the population do not understand the basics of interest compounding and that financially less literate debtors are much more likely to engage in high-cost credit card borrowing. Lusardi and Tufano (2015) find that households with less financial literacy are more frequently unsure about the appropriateness of their debt position and Disney and Gathergood (2013) document that financially illiterate debtors as well as debtors with self-control problems are more likely over-indebted and more frequently fail to repay their debt. Investigating the relationship of financial literacy, gender and credit card behavior, Mottola (2013) finds that women engage more often in costly credit card behavior than men, but that much of the difference can be attributed to demographic characteristics and financial literacy. Lusardi and Tufano (2015) show that women more often rely on high cost borrowing.

However, all existing studies ignore potential gender specific differences in the self-assessment of financial capabilities. Bucher-Koenen et al. (2016) specifically elaborate on the role of women's financial literacy and, in line with the literature, find them to possess severely less financially literacy compared to men. Besides this finding, the authors show women to more often answer financial literacy questions with "do not know" and, when asked to self-assess their financial knowledge, to assign themselves lower scores compared to men. Consequently, the

authors argue that women might be aware of their financial illiteracy or at least unsecure about their financial capabilities. A notion which gains support by Lusardi and Tufano (2015), who find that women, when asked to self-assess the appropriateness of their debt-levels, more frequently answer with "just do not know" compared to men. Given that financially illiterate women might possess a higher awareness of their financial illiteracy or at least seem to be more frequently unsecure about their financial capabilities, we hypothesize that they might restrain themselves from participating in the debt markets in the first place.

Our contribution to the literature is threefold. First, although we find women to possess less financial literacy compared to men on population level, the level of financial literacy for the subsample of debtors does not differ with respect to gender. Thus, our results indicate that financially illiterate women restrain themselves from participating in the debt markets, whereas we cannot observe a similar selection process with respect to financial literacy for men. Second, we show that women are actually better in coping with their debt as they are significantly less often defaulting on their debt, particularly when using unsecured consumer credits which are commonly associated with self-control problems. While the probability of being over-indebted is virtually unchanged for women holding any debt compared to women holding only unsecured consumer debt, the respective probability increases dramatically for men. Finally, for both genders, we find that financial literacy reduces over-indebtedness significantly and we show this effect to be robust against potential endogeneity.

2. Data and methodology

We analyze the determinants of household over-indebtedness using the Panel on Household Finances (PHF), a representative survey of German households by the Deutsche Bundesbank (Deutsche Bundesbank, 2013). The PHF features a rich set of items related to the household balance sheet as well as broad socio-demographic characteristics, allowing profound insights into household's assets and liabilities. The PHF was conducted between September 2010 and July 2011 and includes the responses of 3,565 households.

For the dependent variable in our regressions, we follow Gathergood (2012) and measure household over-indebtedness as actual credit repayment struggles. We classify households as over-indebted if they were unable to make all the due payments on their loans within 12 months before the survey took place. We grasp the households' financial literacy via the three commonly used financial literacy questions introduced by Lusardi and Mitchell (2008), whereby we refer to the sum of correct answers as our measure for financial literacy. To control for the

respondents' formal education, we generate dummies for low-, mid- and high-level education following Dick and Jaroszek (2015). We also control for the respondents' general risk attitude, measured on a scale from 0 [highly risk averse] to 10 [very happy to take risks]. Moreover, we control for potential wealth and employment shocks, the respondents' age, marital status, income and wealth. Gathergood (2012) highlights that unsecured consumer credit, which is frequently used by debtors to facilitate impulse-driven consumption purchases, can be characterized by being easily accessible, comparably costly and having the potential to get out of hand quickly. Women, who are found highly vulnerable to compulsive buying (Achtziger et al., 2015; Dittmar, 2005), might thus be especially endangered to become over-indebted by financing their consumption using unsecured consumer debt. Thus, we acknowledge the distinct characteristics and demands of unsecured consumer debt and, next to our analyses on our whole sample of debtors, run subsample analyses on debtors holding only unsecured consumer debt. Furthermore, as recent literature has acknowledged the potential endogeneity of financial literacy, we estimate linear probability instrumental variable models instrumenting financial literacy using generated instruments after Lewbel (2012). For a complete description of our variables, please refer to Appendix III-1. All analysis are survey weighted and representative for German households.

3. Results

3.1. Descriptive statistics

Table III-1 shows descriptive statistics on German households on population level as well as on the subsample of households holding debt, differentiating between men and women. On the population level, we find the respondents average age to be 52.0 years. Women are significantly less willing to accept risks compared to men, which is in line with recent literature (Almenberg & Dreber, 2015; Bannier & Neubert, 2016). Further, they possess significantly less income and wealth compared to men. In line with, for example, Lusardi and Mitchell (2008) or Bucher-Koenen et al. (2016), we find women to possess significantly less financial literacy compared to men on population level. Nevertheless, women take on debt as often as men. Around one third of all men and women take on any debt, and around 13% of all men and women take on only unsecured debt, indicating that women do not per se restrain themselves from the debt markets.

Table III-1: Descriptive statistics

This table shows descriptive statistics. The data are weighted and representative for German households.

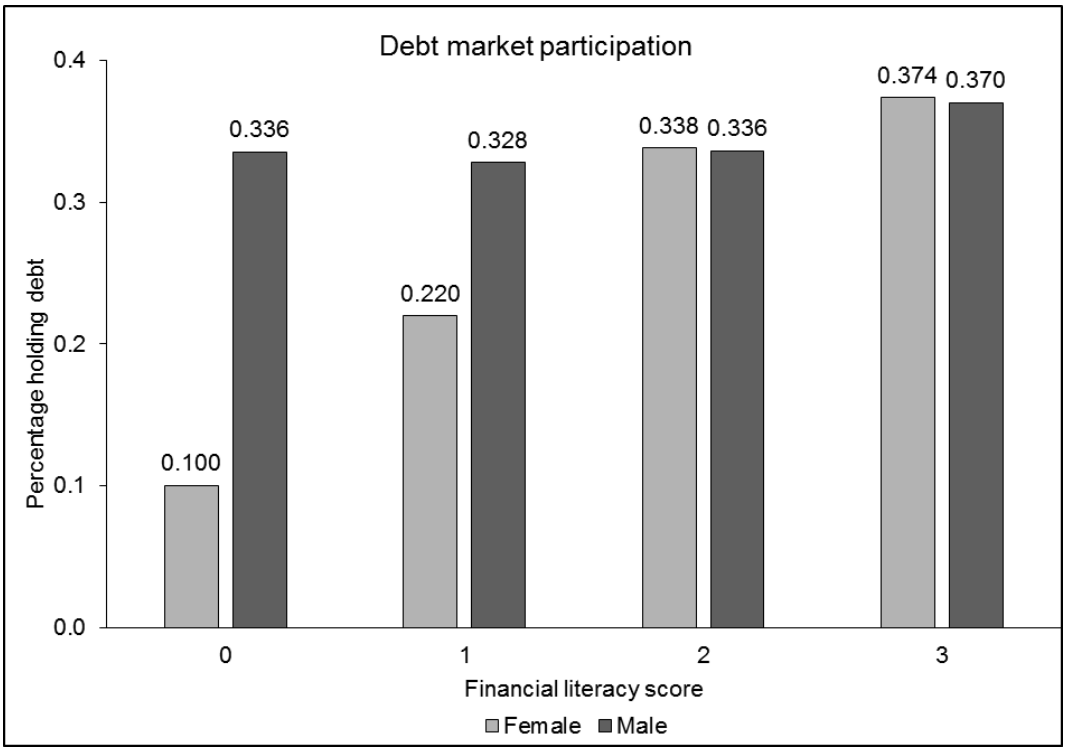
	Debtors					German Population				
Variable	All	Female	Male	Diff	T-Stat	All	Female	Male	Diff	T-Stat
Financial literacy	2.548	2.544	2.551	-0.007	0.12	2.471	2.420	2.518	-0.098	2.52**
Female	0.472					0.490				
Risk attitude	4.0	3.8	4.3	-0.5	2.51**	3.6	3.3	4.0	-0.6	5.47***
Age	46.9	45.4	48.3	-2.9	2.76***	51.2	52.4	51.7	0.7	0.74
Married	0.636	0.613	0.656	-0.044	1.04	0.502	0.456	0.547	-0.091	3.69***
Divorced	0.129	0.116	0.140	-0.025	0.84	0.123	0.143	0.104	0.039	2.27**
Education (low)	0.328	0.261	0.387	-0.126	3.27***	0.419	0.400	0.436	-0.036	1.45
Education (mid)	0.358	0.433	0.291	0.142	3.75***	0.285	0.325	0.246	0.079	3.64***
Education (hi)	0.314	0.306	0.322	-0.016	0.44	0.296	0.275	0.317	-0.043	1.95*
Income	2,983	2,846	3,104	-258	1.30	2,326	2,206	2,441	-235	2.46**
Wealth	193,254	178,240	206,677	-28,437	0.99	156,453	141,739	170,565	-28,826	1.86*
Shock: wealth	0.148	0.180	0.119	0.061	2.09**	0.147	0.168	0.128	0.040	2.30**
Shock: job	0.045	0.065	0.027	0.038	1.99**	0.036	0.042	0.030	0.012	1.26
Debtor						0.344	0.331	0.355	-0.024	1.05
Unsecured debtor	0.365	0.390	0.343	0.047	1.16	0.125	0.129	0.122	0.007	0.43
Debt	75,967	70,892	80,504	-9,611	1.37					
Unsecured debt	13,289	14,855	11,695	3,161	0.68					
Observations	1,381	606	755			3,565	1,596	1,969		

With respect to our sub-sample of respondents holding debt, the average debtor's age is 46.9 years and debt-holding women are 2.9 years younger than men. 47.2% of the debtors are women and, on average, German households owe €75,967 (all debt). Here, women's debt holdings are not statistically different from men's. 36.5% of the debtors in our sample possess only unsecured consumer debt and the respective average amount owed is €13,289. As for debt-holding in general, women do neither differ in their propensity to hold unsecured debt, nor do they hold more unsecured debt in absolute terms compared to men. With respect to formal education, our sample of debtors is quite evenly divided. 32.8% of our households possess only low education, 35.8% mid-level education, and 31.4% higher education. Here, women equally often possess higher education compared to men. With respect to low- and mid-level education, debt-holding women - as opposed to men - possess less frequently low-level education and more frequently mid-level education. Looking at the debt-holding households' risk attitude, women are significantly less willing to accept risks compared to men. Debt-holding women do neither earn significantly less income, nor possess significantly less wealth compared to men.

Looking at the debtors' financial literacy, the average score is 2.55, whereby, in contrast to our findings on population level, female debtholders achieve similar

scores compared to their male counterparts. Given that women as often take on debt as men, the vanished financial literacy gap is quite surprising. Thus, Figure III-1 relates the decision to take on debt to the debtors' gender and financial literacy.

Figure III-1: Debt market participation, financial literacy and gender

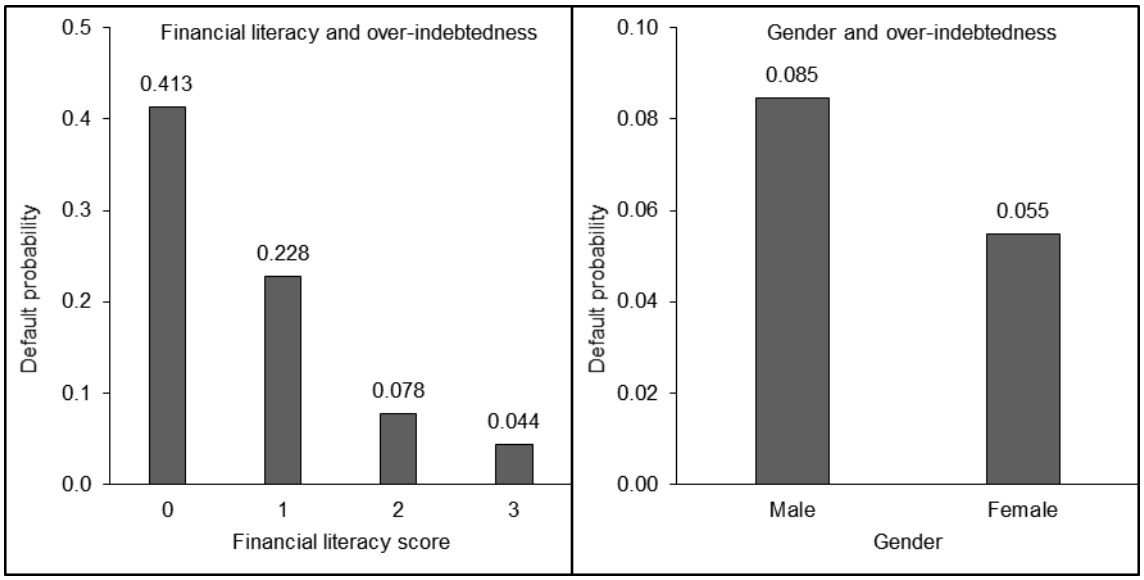


Interestingly, the decision to take on debt seems to be unrelated to financial literacy levels for men. For any level of financial literacy, about one third of the men takes on debt. In contrast, Figure III-1 shows that women with low financial literacy scores less frequently take on debt. Particularly, while the share of debt-holding women equals 37.4% and 33.8% for financial literacy scores of three and two, the respective shares decrease to 22.0% and even 10.0% for women with a financial literacy score of one and zero, respectively. Our results indicate that financially illiterate women, in contrast to their male counterparts, restrain themselves from the debt markets which might be explained by those women being aware of their financial illiteracy or at least being unsecure about their capabilities as argued by Bucher-Koenen et al. (2016).

Figure III-2 relates over-indebtedness to our main explanatory variables, financial literacy and gender, for the sub-sample of debtors. It shows that household

over-indebtedness decreases sharply with each financial literacy question answered correctly. While the average probability for being over-indebted in our sample equals 7.06%, it equals 41.3% for debtors with the lowest possible financial literacy score compared to 4.4% for debtors with the highest scores. With respect to gender, men's probability to be over-indebted equals 8.5%, while it equals only 5.5% for women.

Figure III-2: Over-indebtedness, financial literacy and gender



3.2. Regression Results

Table III-2 reports results from multivariate Probit regressions of financial literacy, gender and a comprehensive set of control variables on over-indebtedness. In an additional setting, we restrict our sample to debtors who hold only unsecured consumer debt. We find financial literacy to be negatively related to over-indebtedness in both settings. Particularly, a one unit increase in the financial literacy score results in a significant reduction of the probability to be over-indebted by 3.2%. For the sample of households holding consumer debt, this effect is more pronounced and equals 4.2%. For our sample of all debtors, being female reduces the probability to suffer from over-indebtedness significantly by 3.3%. When it comes to unsecured consumer debt, being female reduces the probability to be over-indebted by an impressive 10.1%. As univariately found, women seem to be better in coping with their debt-burdens, particularly when it comes to unsecured consumer debt.

Table III-2: Probit regressions

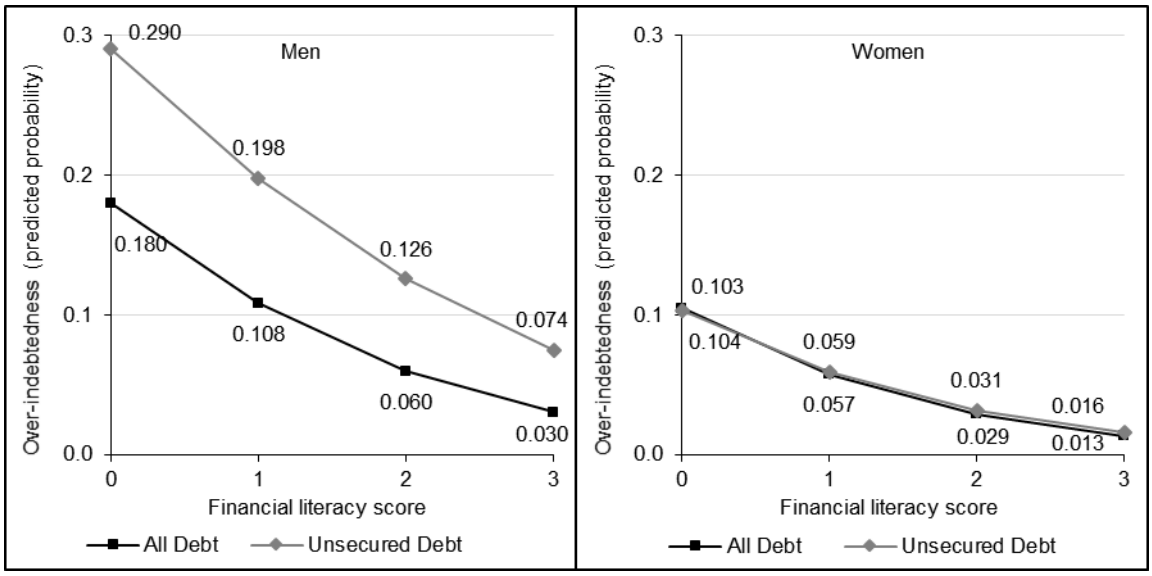
This table shows probit regressions using household over-indebtedness as dependent variable. Tailor linearized standard errors are in parentheses. The data are weighted and representative for German households.
*** p<0.01, ** p<0.05, * p<0.1

	All Debt		Unsecured debt	
	Av. marginal effect	Std. error	Av. marginal effect	Std. error
Financial literacy	-0.032***	(0.011)	-0.042**	(0.021)
Female	-0.033*	(0.018)	-0.101**	(0.041)
Education (mid)	-0.032	(0.020)	-0.056	(0.040)
Education (hi)	-0.024	(0.022)	0.036	(0.057)
Risk attitude	0.007*	(0.004)	0.011	(0.009)
Married	-0.048*	(0.028)	-0.043	(0.043)
Divorced	-0.037*	(0.022)	0.018	(0.056)
Age	0.010*	(0.005)	0.011	(0.012)
Age^2	-0.000*	(0.000)	-0.000	(0.000)
log(income)	-0.045**	(0.019)	-0.097**	(0.046)
log(wealth)	-0.015***	(0.004)	-0.021**	(0.009)
Shock: wealth	0.037	(0.025)	0.093*	(0.050)
Shock: job	0.081	(0.055)	0.187**	(0.084)
log(all debt)	0.006	(0.006)		
log(unsecured debt)			0.015	(0.014)
Observations	1,362		353	
F-test	4.575		2.991	
(p-value)	(0.000)		(0.000)	

In order to further illustrate the effects of financial literacy on over-indebtedness, we present margins plots featuring the predictive probabilities of being over-indebted with respect to gender and debt-type. Holding all remaining explanatory variables at the sample mean, Figure III-3 shows that the households' probability to be over-indebted declines sharply with the financial literacy score for both genders. We find the probabilities of being over-indebted to be generally higher for men compared to women. For all debt, the probability of being over-indebted for a man who answers no financial literacy question correctly equals 18.0% and declines to 3.0% if he answers all three questions correctly. For women, in contrast, the respective probabilities are with 10.4% and 1.3% severely lower. In particular, a debt-holding man with a financial literacy score of zero is 1.7 times, and a man with a financial literacy score of three is 2.3 times more likely to be over-indebted when compared to a woman with the same financial literacy scores. Looking at unsecured consumer debt, the respective man's probabilities to be over-indebted dramatically increase. Here, the probability of being over-indebted for a man who answers no financial literacy question correctly equals 29.0% and still remains 7.4% if he answers all three questions correctly. The respective probabilities for a woman holding unsecured consumer debt are rather unchanged

compared to a woman holding any debt and equal 10.3% and 1.6%, respectively. Hence, a man with a financial literacy score of zero holding unsecured debt is 2.8 times, and a man with a financial literacy score of three is 4.6 times more likely to be over-indebted when compared to a woman with the same financial literacy scores. As unsecured consumer debt are frequently used to facilitate impulse-driven consumption purchases and thus require a higher amount of self-control (Gathergood, 2012), our results not only indicate that women are better in coping with their debt burdens, but also that they, despite being commonly found to be highly vulnerable to compulsive buying (Achtziger et al., 2015; Dittmar, 2005), seem to be less prone to the self-control problems associated with unsecured consumer debt.

Figure III-3: Marginal effects from Probit models in Table III-2



We control our regression results for potential endogeneity of financial literacy by estimating linear probability instrumental variable models instrumenting financial literacy using generated instruments after Lewbel (2012). Table III-3 shows that the instrumented coefficients remain economically and statistically significant and the endogeneity tests indicate that the null hypothesis of the financial literacy score's exogeneity cannot be rejected.

Table III-3: IV Regressions with generated instruments

This table presents second stage IV GMM linear probability model estimates of our baseline models in Table III-2, instrumenting financial literacy using generated instruments after Lewbel (2012). Standard errors are robust. All data are weighted and representative for German households.

	All Debt		Unsecured Debt	
	Marginal Effect	Std. Error	Marginal Effect	Std. Error
Financial literacy	-0.040*	(0.024)	-0.070*	(0.039)
Female	-0.025	(0.019)	-0.098**	(0.042)
Education (mid)	-0.031	(0.026)	-0.035	(0.048)
Education (hi)	-0.030	(0.026)	0.016	(0.052)
Risk attitude	0.009**	(0.004)	0.010	(0.010)
Married	-0.044	(0.032)	-0.027	(0.042)
Divorced	-0.063	(0.046)	0.016	(0.074)
Age	0.006	(0.005)	0.005	(0.009)
Age^2	-0.000	(0.000)	-0.000	(0.000)
log(income)	-0.032*	(0.017)	-0.091*	(0.047)
log(wealth)	-0.017***	(0.005)	-0.013	(0.008)
Shock: wealth	0.033	(0.032)	0.083	(0.059)
Shock: job	0.041	(0.056)	0.187**	(0.085)
log(all debt)	0.009	(0.007)		
log(unsecured debt)			0.020	(0.012)
Constant	0.374**	(0.164)	0.802**	(0.354)
Observations	1,362		353	
R-squared	0.12		0.21	
Endog test	1.14		1.16	
(p-value)	(0.29)		(0.28)	
Hansen J statistic	14.46		17.03	
(p-value)	(0.27)		(0.15)	
F-Test of excluded Instruments	53.94		22.86	
(p-value)	(0.00)		(0.00)	

4. Conclusion

Our results highlight the importance of a sufficient understanding of basic financial concepts in order to cope with everyday financial decisions. In line with the literature, we find women to possess less financial literacy compared to men on population level. However, when it comes to holding debt, only financially literate women participate in the debt markets, indicating that financially illiterate women might be aware of their financial illiteracy or at least seem to be unsecure about their financial capabilities and restrain themselves from debt markets. Those women who hold debt seem to be better able in coping with their debt burdens compared to men, as we find them to be less often over-indebted. Generally, but particularly when it comes to unsecured consumer debt, men are dramatically more often over-indebted than women. Thus, our results might in-

dicate that women, who are commonly found to be highly vulnerable to compulsive buying, are less susceptible for such self-control problems when it comes to their financing decisions. Finally, using a robust measure of over-indebtedness, we find that financial literacy reduces over-indebtedness significantly for both genders and we show this effect to be robust against potential endogeneity.

5. References

Achtziger, A., Hubert, M., Kenning, P., Raab, G., & Reisch, L. (2015). Debt out of control: The links between self-control, compulsive buying, and real debts. *Journal of Economic Psychology*, 49, 141-149.

Almenberg, J., & Dreber, A. (2015). Gender, stock market participation and financial literacy. *Economics Letters*, 137, 140-142.

Bannier, C. E., & Neubert, M. (2016). Gender differences in financial risk taking: The role of financial literacy and risk tolerance. *Economics Letters*, 145, 130-135.

Bucher-Koenen, T., & Lusardi, A. (2011). Financial literacy and retirement planning in Germany. *Journal of Pension Economics and Finance*, 10(04), 565-584.

Bucher-Koenen, T., Lusardi, A., Alessie, R., & van Rooij, M. (2016). How financially literate are women? An overview and new insights. *Journal of Consumer Affairs*. doi:10.1111/joca.12121

Deutsche Bundesbank. (2013). Panel on household finances. doi:10.12757/PHF.01.01.01.stata

Dick, C. D., & Jaroszek, L. (2015). Think twice or be wise in consumer credit choices. *Working Paper*.

Disney, R., & Gathergood, J. (2013). Financial literacy and consumer credit portfolios. *Journal of Banking & Finance*, 37(7), 2246-2254.

Dittmar, H. (2005). Compulsive buying - a growing concern? An examination of gender, age, and endorsement of materialistic values as predictors. *British journal of psychology* (London, England: 1953), 96(Pt 4), 467-491.

Gathergood, J. (2012). Self-control, financial literacy and consumer over-indebtedness. *Journal of Economic Psychology*, 33(3), 590-602.

Lewbel, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics*, 30(1), 67-80.

Lusardi, A., & Mitchell, O. S. (2008). Planning and financial literacy how do women fare? *American Economic Review*, 98(2), 413-417.

Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy theory and evidence. *Journal of Economic Literature*, 52(1), 5-44.

Lusardi, A., & Scheresberg, C. d. B. (2013). Financial literacy and high-cost borrowing in the United States. *Working Paper*.

Lusardi, A., & Tufano, P. (2015). Debt literacy, financial experiences, and overindebtedness. *Journal of Pension Economics and Finance*, 14(04), 332-368.

Mottola, G. (2013). In our best interest: Women, financial literacy, and credit card behavior. *Numeracy*. doi:10.5038/1936-4660.6.2.4

Stolper, O. A., & Walter, A. (2017). Financial literacy, financial advice, and financial behavior. *Journal of Business Economics*.

van Rooij, M. C., Lusardi, A., & Alessie, R. J. (2011). Financial literacy and stock market participation. *Journal of Financial Economics*, 101(2), 449-472.

von Gaudecker, H.-M. (2015). How does household portfolio diversification vary with financial literacy and financial advice? *The Journal of Finance*, 70(2), 489-507.

.

6. Appendix

Appendix III-1: Variable descriptions

This table contents the descriptions of our variables in alphabetical order.

Variable Name	Variable Description
Age	Continuous variable that measures the respondent's age in (years).
Over-indebted	Dummy variable that equals one if the respondent was not able pay a monthly installment within the last 12 months. Zero otherwise. Underlying PHF Item: (Were you / Was your household / Was the household) able to make all the due payments for the various loans, mortgage loans and leasing contracts on time over the past 12 months?
Divorced	Dummy variable that equals one if the respondent is divorced. Zero otherwise.
Debt (all)	Continuous variable that measures the household's amount of debt (EUR).
Debt (unsecured)	Continuous variable that measures the household's amount of unsecured debt (EUR).
Education (low)	Dummy variable that equals one if the respondent has low-level education ("Hauptschulabschluss" or lower), zero otherwise.
Education (mid)	Dummy variable that equals one if the respondent has mid-level education ("Mittlere Reife"), zero otherwise.
Education (high)	Dummy variable that equals one if the respondent has A-level education ("(Fach-)Hochschulreife"), zero otherwise.
Female	Dummy variable that equals one if the respondent is female, zero otherwise.
Financial Literacy	Ordinal variable that measures the number of correctly answered financial literacy questions. Underlying PHF-items: *Question 1: Compound Interest effect* Let us assume that you have a balance of €100 on your savings account. This balance bears interest at a rate of 2% per year and you leave it for 5 years on this account. How high do you think your balance will be after 5 years? 1 - More than €102 [correct] 2 - Exactly €102 3 - Less than €102 *Question 2: Inflation* Let us assume that your savings account bears interest at a rate of 1% per year and the rate of inflation is 2% per year. Do you think that in one year's time the balance on your savings account will buy the same as, more than or less than today 1 - More 2 - The same 3 - Less than today [correct] *Question 3: Diversification* Do you agree with the following statement: "Investing in shares of one company is less risky than investing in a fund containing shares of similar companies"? 1 - Agree 2 – Disagree [correct]
Income	Continuous variable that measures the household's monthly income in (EUR).
Shock: wealth	Dummy variable that equals one if the respondent's net worth has decreased substantially during the last two years. Zero otherwise.
Shock: job	Dummy variable that equals one if the respondent lost his/her job within the last two years. Zero otherwise.
Wealth	Continuous variable that measures the household's gross wealth in (EUR).
Married	Dummy variable that equals one if the respondent is married, zero otherwise.
Risk Attitude	Ordinal variable that measures the respondents' propensity to take financial risks on a scale from 0 [highly risk-averse] to 10 [very happy to take risks].

IV. Content analysis of business-specific text documents: Introducing a German dictionary

Co-authors: Christina E. Bannier, Andreas Walter
Own share: 80%

This paper was presented on the following refereed conferences/ workshops:

- 79th Annual Meeting of the German Academic Association for Business Research (VHB), St. Gallen, Switzerland, 2017.*
- International Conference on Discourse approaches to financial communication (DAFC), Lugano, Switzerland, 2017.*
- PhD Workshop of the German Finance Association (DGF), Bonn, Germany, 2016.

(* denotes forthcoming presentations)

This paper was presented on the following non-refereed conferences/ workshops:

- GGS doctoral seminar, Gießen, Germany, 2016.

Content analysis of business-specific text documents: Introducing a German dictionary

Christina E. Bannier[a] Thomas Pauls[b] Andreas Walter[c]

Abstract - Computer-aided quantitative content analyses have recently gained a lot of attention. Applied on different elements of business communication such as financial disclosures, analyst reports, earnings announcements or IPO prospectuses, they have been shown to deliver relevant information to financial market participants. However, most analyses of business-specific texts have been conducted using English documents solely. To contribute to a wider usage of the analytical instruments developed so far, we create a content-specific German dictionary for textual sentiment analysis. The so-called BPW dictionary is based on the widely used English language dictionary by Loughran and McDonald (2011). We extensively evaluate our dictionary and find our adaptation to be widely equivalent compared to its English original.

Keywords: Textual sentiment, Market efficiency, Textual analysis, Annual reports

JEL-Codes: G02, G12, G14

[a] Department of Banking & Finance, University of Gießen, Licher Str. 62, 35394 Gießen, Germany. Christina.Bannier@wirtschaft.uni-giessen.de.

[b] Department of Financial Services, University of Gießen, Licher Str. 74, 35394 Gießen, Germany. Thomas.Pauls@wirtschaft.uni-giessen.de.

[c] Department of Financial Services, University of Gießen, Licher Str. 74, 35394 Gießen, Germany. Andreas.Walter@wirtschaft.uni-giessen.de.

1. Introduction

Recent financial research has acknowledged the information value of the qualitative characteristics of text documents, i.e., their textual sentiment (Kearney & Liu, 2014; Loughran & McDonald, 2016). Fostered by an increasing computing performance, the dictionary-based approach has become a commonly used tool to capture the sentiment of various kinds of documents such as, e.g., financial disclosures, analyst reports, earnings press releases, IPO prospectuses, internet board postings, or newspaper articles (Kearney & Liu, 2014). Particularly with regard to finance-related applications, Loughran and McDonald (2015) have shown the superiority of context-specific dictionaries over the use of more general word lists. However, hardly any adaptations of business-specific dictionaries exist outside the English language. This may come as a surprise since many text documents containing value-relevant information like news articles, product information, and business communication, are published only in the native language. Even if these documents are translated into English, they are typically altered, shortened, or summarized in the translation process, so that the original sentiment gets obscured. In addition, the translated versions are often published later than the initial documents, so that it will be hard to profit from an analysis of their sentiment on financial markets.

We try to fill this gap and contribute to broadening the use of dictionary-based content analyses by creating a business-specific dictionary for the German language. Following the methodology of Wolf et al. (2008), we translate and adjust the positive, negative, uncertainty wordlists of the dictionary compiled by Loughran and McDonald (2011), which has evolved as the industry standard in textual analysis of documents in the fields of finance and accounting. We refer to our German adaptation as the BPW dictionary. We test the equivalence of our adaptation using a sample of 1,402 quarterly and annual reports of German DAX and MDAX companies which are available in German and English. For that purpose, we estimate simple pairwise correlations, Spearman rank correlations, intraclass correlations (Shrout & Fleiss, 1979), and test the dictionaries' equivalence via two-sided equivalence testing as introduced by Blair and Cole (2002). Our results show a high correlation and equivalence to the English counterpart, indicating the reliability of our adaptation. Finally, we compare our dictionary to two general German dictionaries and present evidence for the BPW dictionary being better suited to capture the sentiment of business-specific documents. Our contribution to the literature is threefold. First, we create a tool that allows a context-specific evaluation of the qualitative information disclosed in German text documents. It broadens the applicability of content analyses as our dictionary is

not sample specific, i.e., it can be used on any type of business communication with a focus on delivering value-relevant information. The BPW dictionary will be available at www.uni-giessen.de/BPW, paving the way for further German content analyses on business related documents. Second, we compare our context-specific dictionary with general language dictionaries and, in line with studies on the English language, find evidence for the superiority of context-specific dictionaries in measuring the textual sentiment of documents. Finally, by giving a detailed account of the adaptation process, we describe a framework for future adaptations of dictionaries in other languages.

The paper is structured as follows. The next section reviews the literature on content analyses. Section 3 describes the underlying dictionary introduced by Loughran and McDonald (2011) as well as the adaptation process for creating the BPW dictionary. In section 4, we present our adaptation, show the results of the correlation and equivalence tests and compare our dictionary to general German language dictionaries. Section 5 concludes.

2. Literature

2.1. General language dictionaries

Early studies conducting content analyses on business-related documents utilize the comprehensive Harvard University's General Inquirer IV-4[1] (hereafter HARVARD) dictionary introduced by Stone et al. (1966). HARVARD contains dozens of categories, such as, positive, negative, strong, weak, political, or religious to capture the sentiment of a text. Its positive and negative wordlists contain 2,577 and 3,699 words, respectively. Despite its general character, the HARVARD dictionary has been used frequently in the finance literature. For example, Tetlock (2007) and Tetlock et al. (2008) use the HARVARD dictionary and measure the pessimism in news articles. They find that high values of media pessimism convey negative information about firms' earnings and induce downward pressure on market prices. The HARVARD dictionary is also used for analyzing text documents of 10-Ks and 10-Qs filings, earnings conference calls, earnings press releases, and news articles (Kearney & Liu, 2014).

Another general language dictionary, which was repeatedly used in the financial literature, is included in the Diction[2] software (hereafter DICTION). Its original area of application is the analysis of political language and rhetorical analysis.

[1] See http://www.wjh.harvard.edu/~inquirer/.

[2] See http://www.dictionsoftware.com/.

In its latest version, DICTION comprises 10,000 words and 35 sentimental categories, of which the negative wordlist includes 914 words and the positive wordlist 697. Davis et al. (2012) use the DICTION dictionary to gauge the optimism in earnings press releases and find it to be predictive of future firm performance. In a subsequent study, Davis and Tama-Sweet (2012) employ the DICTION dictionary to investigate the sentiment of earnings press releases and of the corresponding management discussion and analyses (MD&A) sections in the companies' annual reports. The authors find that firms which exactly meet or just beat earnings benchmarks report a lower proportion of total pessimistic language in their earnings press releases. In addition, they find the level of pessimistic language in the MD&A sections to be predictive of future firm performance. Furthermore, DICTION was also used to examine earnings conference calls (Davis et al., 2012) and IPO prospectuses (Ferris et al., 2013).

A third general language dictionary is the Linguistic Inquiry Word Count[3]. Its purpose is to capture people's social and psychological states. Its language categories are created by researchers with a special interest in social, clinical, health, and cognitive psychology. The Linguistic Inquiry Word Count contains 72 sentimental categories and more than 2,300 different words. It has been used in numerous studies containing social-, personality-, or clinical psychological research questions on a variety of different text-documents, such as lyric, therapeutic essays, political speeches, daily conversations, and computer-based communication (Wolf et al., 2008). Larcker and Zakolyukina (2012) utilize the Linguistic Inquiry Word Count to gauge deceptive tone of CEO and CFO narratives during earnings conference calls. The authors show the textual sentiment to be indicative of financial misreporting. However, Larcker and Zakolyukina (2012) admit that the Linguistic Inquiry Word Count as a general language dictionary might not be appropriate for capturing the sentiment with a business-specific focus. According to Loughran and McDonald (2015) this criticism also applies to both the HARVARD and DICTION dictionaries.

2.2. Context-specific dictionaries

Henry (2008) was the first to compose a dictionary explicitly designed to examine the tone of financial documents. She customizes a dictionary on a specific financial text sample: earnings press releases. Her dictionary reveals that the earnings press releases' positive sentiment positively affects the market reaction subsequent to the earnings announcements. Due to the customization to one specific

[3] See http://www.liwc.net.

text type, the Henry (2008) (hereafter HENRY) dictionary contains only 85 negative and 105 positive words. Despite the small number of words, various studies describe the superiority of the context-specific HENRY dictionary over the DICTION and HARVARD dictionaries (Doran et al., 2012; Henry & Leone, 2016; Price et al., 2012) when applied to business-related text documents.

Nonetheless, the HENRY dictionary's applicability is clearly limited by its small number of words included. Hence, Loughran and McDonald (2011) create a comprehensive dictionary by evaluating all words appearing in at least 5% of the entire 10-K disclosure universe. Apart from a positive and negative wordlist to capture optimism and pessimism in a text, it also contains wordlists for uncertainty or modal words to assess a text's "tonality" in a broader sense. The Loughran and McDonald (2011) (hereafter LM) dictionary contains 2,354 negative and 354 positive words. The authors state that 73.8% of the HARVARD negative word count does actually not have a negative meaning in financial documents. In subsequent work, the authors compare their dictionary with the HENRY dictionary and emphasize that none of the most frequently occurring negative words of the LM dictionary (loss, losses, claims, impairment, against, adverse, restated, adversely, restructuring, and litigation) are included in the HENRY dictionary (Loughran & McDonald, 2015). In addition to that, Loughran and McDonald (2015) use the LM and the DICTION dictionaries to gauge the sentiment of financial disclosures and compare both dictionaries' explanatory power on post 10-k filing stock return volatility. The authors find 83% of the words in the DICTIONs positive wordlist and 70% of the words in the DICTION negative wordlist to be misclassified. Furthermore, the sentiment gauged by the LM dictionary seems to significantly better explain post 10-k filing stock return volatility. Thus, the authors argue that DICTION, as a general dictionary, is inappropriate to assess the sentiment of financial documents. Due to its comprehensiveness and its appropriateness for financial documents, the LM dictionary has become the most widely used dictionary in business research. It has been utilized to assess the tone of a variety of different documents, for example, 10-k filings (Loughran & McDonald, 2011), earnings conference calls (Davis et al., 2015), news articles (García, 2013), and IPO prospectuses (Ferris et al., 2013; Jegadeesh & Wu, 2013).[4]

[4] For a comprehensive overview over studies using mentioned general and context-specific dictionaries, see Kearney and Liu (2014) and Loughran and McDonald (2016).

2.3. German dictionaries

When it comes to the analysis of German text documents, two comprehensive general German language dictionaries exist. Remus et al. (2010) created the "SentimentWortschatz" (hereafter SENTIWS) dictionary, which is based on and extends the General Inquirer lexicon by Stone et al. (1966). The SENTIWS dictionary comprises 15,466 negative and 15,536 positive individual words and has been used in the fields of, for example, political communication (Haselmayer & Jenny, 2016), as well as art and literature (Zehe et al., 2016). The second general language dictionary was created by Wolf et al. (2008), who adapted the English version of the Linguistic Inquiry Word Count to the German language. Their dictionary (hereafter LIWC) contains 1,049 negative and 646 positive words and, like its English original, puts special emphasis on analyses of essays in the context of expressive writing experiments. It has also been used in other research domains such as, for example, political analysis (Caton et al., 2015; Jacobi et al., 2016).

However, no comprehensive context-specific dictionary for the analysis of business-related text-documents exists in the German language. This is despite the notion that, like for the English language, German general language dictionaries are likely to be inferior compared to context-specific dictionaries in assessing the textual sentiment of business-related text documents. In this respect, Ammann and Schaub (2016) analyze data from an online social trading network, where traders publish their trading strategies for followers to comment on and invest in. Using a sample specific ad-hoc dictionary, they find that online investors adjust their trading behavior to the commentaries' sentiment but the commentaries do not seem to have predictive power for the trading strategies' future performance. Next to their ad-hoc dictionary, the authors also utilize the general SENTIWS and LIWC dictionaries and derive similar, albeit weaker results. Mengelkamp et al. (2016) also create simple, sample specific ad-hoc dictionaries from a manually categorized subsample of Twitter messages and find them to perform quite well. Furthermore, the authors compare their ad-hoc dictionaries' performance to that of SENTIWS and report a significantly higher performance of their specific dictionaries. Thus, Ammann and Schaub (2016) as well as Mengelkamp et al. (2016) highlight the superiority of content-specific dictionaries over general language dictionaries. However, their ad-hoc dictionaries are restrictive in their use of specific sample text documents. Deriving a sample-independent context-specific dictionary would therefore be clearly preferable.

3. The creation of the BPW dictionary

3.1. Translating a dictionary

Wolf et al. (2008) adapt the English version of the Linguistic Inquiry Word Count dictionary to the German language. Their methodology follows the translation of the Linguistic Inquiry Word Count into other languages such as, for example, Dutch (Zijlstra et al., 2004) or Spanish (Ramírez-Esparza et al., 2007).[5] For the translation, all words in each category of the English version are summarized in a table and, with the aid of common English-German dictionaries, a parallel version is created and supplemented by meaningful and relevant related words. 25% of the words are retranslated to check their meaning equivalence. Afterwards, Wolf et al. (2008) utilize 122 English text-documents and their respective German translations and estimate a two-sided equivalence test of the difference between the means following Blair and Cole (2002) as well as the intraclass correlations (ICC) following Shrout and Fleiss (1979) in order to test their translated German dictionary's equivalence to the English version. For the adaption of the LM dictionary, we build upon the methodology of Wolf et al. (2008), using 1,402 German quarterly and annual reports and their corresponding English versions for the equivalence tests. However, we adjust their methodology by accounting for a number of linguistic issues as explained in the following.

3.2. Inflections

Probably the most prevalent issue to be considered in the adaptation process are differences in the inflectional morphology between German and English. In general, there are two possibilities to account for inflections using the dictionary-based approach: One could include all possible inflections of a word into the dictionary. Alternatively, one could reduce morphological variants of words to their word stem or root form (stemming). Unlike the Linguistic Inquiry Word Count, the LM dictionary uses inflections to avoid errors associated with stemming. Loughran and McDonald (2011) draw attention to the problem that a word's meaning might change when common prefixes or suffixes are added. For example, "ODD" and "BITTER" take on different meanings when made plural: "ODDS" and "BITTERS". As a consequence, the authors advise using explicit inflections as these are less prone to error than using stemming (Loughran & McDonald, 2015). This is true also for the German language. For example, the negative wordlist of the German version of the Linguistic Inquiry Word Count includes the stem

[5] The LIWC dictionaries are also translated in Norwegian, Italian, and Portuguese. Other translations into Arabic, Korean, Turkish, and Chinese are in progress (Pennebaker et al., 2015).

"WEINE", which is the stemmed form of the verb "WEINEN" (in English to cry). However, "WEINE" is also the plural form of "WEIN", which is the German word for wine.

As German grammar is more explicit compared to English (Hawkins, 2015; König & Gast, 2012), the use of inflections creates specific problems when it comes to German language. Looking at nouns, both languages distinguish with respect to singular and plural. However, the German language further distinguishes four cases in the noun phrase: nominative, accusative, genitive, and dative. In contrast, English only retains a separate genitive case which is not relevant in terms of computational analysis as the English genitive form is reduced to the respective singular or plural stem in the tokenization process (Hawkins, 2015). With respect to verbs, German distinguishes indicative and subjunctive forms whereas English employs a single form for both. Further, German verbs are distinguished with respect to person and number, whereas the bare stem in English is used for all persons and numbers except for the third person singular (Hawkins, 2015). For example, the English present tense forms "LAY" and "LAYS" correspond to the German indicatives "LEGE", "LEGST", "LEGT", "LEGEN", and "LEGET", while the past tense form "LAID" corresponds to the German "LEGTE", "LEGTEST", "LEGTE", "LEGTEN", and "LEGTET" (König & Gast, 2012). Thus, less inflections are needed to comprise the morphology of English adjectives. Likewise, the only inflectional forms we find for English adjectives are the comparative and superlative forms, for example, "HAPPY", "HAPPIER", and "HAPPIEST". Further, when constructed with a verb or another adjective, English adjectives attach the marker "–ly" (i.e. "HAPPILY") (König & Gast, 2012). In German, adjectives are distinguished with respect to gender, case (nominative, accusative, dative, and genitive), as well as comparative and superlative forms. Consequently, the four inflections of "happy" correspond to the German "GLÜCKLICH", "GLÜCKLICHE", "GLÜCKLICHER", "GLÜCKLICHEN", "GLÜCKLICHEM", "GLÜCKLICHES", the comparative forms "GLÜCKLICHER", "GLÜCKLICHERE" and "GLÜCKLICHERES", and the superlative forms "GLÜCKLICHSTE", "GLÜCKLICHSTES", "GLÜCKLICHSTEN", "GLÜCKLICHSTEM" (König & Gast, 2012).

In order to address the issue of inflectional morphology, we therefore systematically generate German inflections next to the direct translations and add them to our wordlists. However, a certain inflection might have additional meanings that may not fit into the respective sentimental wordlist. For example, one translation of the English verb "CUT", which is included in the LM negative wordlist, is the German verb "KÜRZEN". The German first person singular is "KÜRZE", which might not only be retranslated into "CUT", but also into "SHORTNESS" or "BRIEFNESS". In the latter cases, the retranslations would not have a negative

context. To address this issue, we retranslate each generated inflection using common German-English dictionaries and exclude inflections that have additional meanings and do not fit into the respective sentimental category of the English original word.

3.3. Lexical morphology

German and English are also different with respect to lexical morphology. In particular, German not only requires more inflectional accuracy, but also more semantic distinctions within a lexical field where English uses undifferentiated and broader terms (Hawkins, 2015). For example, the English verb "STOP" translates into the German verbs "AUFHÖREN", "HALTEN", "STEHENBLEIBEN", "AUFHALTEN", "INNEHALTEN", "ANHALTEN", and "(UNTER-)LASSEN". German speakers have to identify the specific type of "STOP" which is defined by the context in which it is used. For example, "to stop working" translates into "aufhören zu arbeiten", "to stop somebody (from doing something)" translates into "jemanden aufhalten (etwas zu tun)", or "to stop a car" translates into "ein Auto anhalten". Each German verb is restricted in its semantic coverage, whereas the English single verb extends over all the semantic distinctions. In order to address the differences in lexical morphology, we, in a first step, gather all translations using common English – German dictionaries. In a second step, we retranslate each translation and exclude those translations which are not of clear sentimental implication with respect to the sentimental wordlist the English original word was included in. The English verb "ESCALATE", which is included in the LM dictionary's negative wordlist, might serve as an example. Its German translations are "ANSTEIGEN", "AUSUFERN", "ZUSPITZEN", and "ESKALIEREN". While the latter three translations will most likely be used to describe negatively connoted events, "ANSTEIGEN" might be used to describe the growth of, for example, costs, which might be negative in context, but also the growth of, for example, sales, which would be positive in context. Consequently, we add "AUSUFERN", "ZUSPITZEN", and "ESKALIEREN" to the negative wordlist of our dictionary while we do not include "ANSTEIGEN."

3.4. Compound words

German compound words are described by a single word in German whereas they are described by several words in English (König & Gast, 2012). For example, the German noun "LEBENSVERSICHERUNGSFACHANGESTELLTER" translates into "LIFE INSURANCE COMPANY EMPLOYEE". This may easily lead to problems in a content analysis. For example, the LM uncertainty word list includes the word "RISK", whose German translation is "RISIKO". Including

only the direct translation "RISIKO" would underestimate the German share of uncertain words. This is because German compound words such as, for example, "AUSFALLRISIKO" (default risk), "LIQUIDITÄTSRISIKO" (liquidity risk), "MARKTRISIKO" (market risk), "FINANZRISIKO" (financial risk), "KREDITRISIKO" (credit risk), "WÄHRUNGSRISIKO" (currency risk), "ZINSRISIKO" (interest rate risk), "GESCHÄFTSRISIKO" (business risk), or "REPUTATIONSRISIKO" (reputation risk) would then not be assessed as uncertain using the German dictionary. As a consequence, we explicitly search for German compound words and add them to our dictionary.

3.5. Retranslation

In a final step, we retranslate all words in our dictionary, review their meaning with respect to the sentimental category and drop non-fitting words that may have multiple meanings which do not all correspond to the respective sentimental category. Retranslation confirmability, which represents the share of words which remain in our wordlists after first including all possible translations from the LM dictionary and then retranslating all words with special attention to their fit into the respective sentiment category, takes on values of 84.34%, 75.61%, and 87.43% for the negative, positive and uncertainty list, respectively. This is close to the values in Wolf et al. (2008), who retranslate only 25% of their words and observe a retranslation conformability of 89%. Table IV-1 provides the final number of words in our (hereafter BPW) dictionary's wordlists and those of the LM, HENRY, HARVARD and DICTION dictionaries.

Table IV-1: Number of words in wordlists

This table shows the number of words in the English language dictionary's wordlists and the corresponding number of words in our adaption.

	BPW	LM	HENRY	HARVARD	DICTION
Negative	10,147	2,354	85	3,699	914
Positive	2,223	354	105	2,577	697
Uncertainty	1,697	297			

4. Evaluation

4.1. Equivalence of the BPW to the LM dictionary

After creating the BPW dictionary, we evaluate its quality of fit compared to its English counterpart, the LM dictionary. For this evaluation, we utilize quarterly and annual reports from German DAX and MDAX companies from end 2008 to early 2015. We gather those reports from DGAP, a German information

platform where stock-listed companies publish company-specific news in order to fulfil ad-hoc and other obligations.[6] Our sample consists of 1,402 German reports and their corresponding English versions. These extensive reports are publicly available in German and in English for all listed companies in Germany, they are translated by professional translation bureaus, and are released simultaneously.

Before we can analyze the reports quantitatively, we have to convert the documents, which are typically available in PDF file format, to TXT format. Thereby, we replace typographic ligatures and employ UTF-8 character encoding on all files in order to allow for German-specific characters such as 'Ä','Ü','Ö', or 'ß'. All characters are transformed into lower case and tokenized afterwards, where we define a token as any subsequent order of at least three alphabetic characters. In order to exclude potential spelling errors, we exclude tokens that do not occur in at least one percent of the reports.

Afterwards, we apply a stop-word list on the reports in both languages to filter words that might have important semantic functions, but that rarely contribute information (Manning & Schütze, 1999). For the English versions of the reports, we use the stop-word list provided by Loughran and McDonald which includes common names, dates, numbers, geographic locations, or currencies.[7] We supplement the list, amongst others, with the names of German DAX and MDAX companies, popular German pre- and surnames, and the names of the largest German and European cities. For the German reports, we adapt the stop-word list by adding the German translations.[8] The dictionary-based approach, which is thus also called bag-of-words approach, implies words to be independent from each other and dissolves the sequential order of words. Consequently, the documents are hereafter transformed to word count vectors using the Rapidminer software.[9] In a final step, the German and English documents' sentiment is assessed by counting their quantity of negative, positive, and uncertain words with respect to the word lists of the BPW and LM dictionary, respectively. The corresponding shares of sentimental words are then calculated by dividing the number of sentimental words with respect to one category by the respective report's total number of words. For the English versions of the reports, the textual sentiment is hence measured using the LM dictionary's wordlists and for the German versions of the reports, the textual sentiment is measured using our BPW dictionary.

---

[6] For more information, please see http://www.dgap.de/.

[7] For more information, please see http://www3.nd.edu/~mcdonald/Word_Lists.html.

[8] We will provide the stop word list at www.uni-giessen.de/bpw.

[9] The transformation to lower-case characters, the tokenization, the stop-word filtering and the generation of the word count vectors were conducted with the Rapidminer software. For more information, please see https://rapidminer.com/.

Note that the use of quarterly and annual reports obliges make two exemptions from the dictionary-based approach's word independence assumption. Specifically, we have to control for certain combinations of "PROFITS", "GAINS", and "LOSSES" as not doing so would result in an overestimation of the English reports' shares of positive and negative words compared to their German counterparts. See Appendix IV-1 for complete list of all combinations we controlled for in our analyses on our sample of quarterly and annual reports. Further, for the same reason, we counted the terms "IMPAIRMENT LOSS" and "IMPAIRMENT LOSSES" as one negative occurrence. We also conduct our analysis on the quarterly and annual reports without making any exception from the word independence assumption. The results are presented in Appendix IV-2.

Table IV-2 provides summary statistics for our set of quarterly and annual reports. While the German versions of the reports contain only slightly more words in total compared to their English counterparts, they contain significantly more individual words per document: The German reports contain 2.63% (4.90%) more total words in the mean (median) and 66.08% (51.24%) more individual words per report in the mean (median). As the content of the corresponding texts should be broadly identical, the large difference highlights that German has more distinct word forms compared to English and underlines the necessity for the adjustments in our adaptation process.

Table IV-2: Summary statistics of the quarterly and annual reports

This table presents summary statistics of our sample of corresponding 1,402 English and German quarterly and annual reports.

		Total	Words per document					Individual words per document				
	Reports	Words	Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
ENG	1,402	33,226,602	23,699	11,398	26,486	444	185,071	2,432	1,780	1,571	122	7,806
GER	1,402	34,099,075	24,322	11,957	27,143	396	195,713	4,039	2,692	3,054	150	15,237

Table IV-3 reports the ten most common English and German negative, positive, and uncertain words in our sample of quarterly and annual reports according to the LM dictionary and our BPW dictionary respectively. Comparing our results from the English versions of the reports with those of the German versions, the total numbers of sentimental words are quite comparable, providing first evidence for the equivalence of our translation. Moreover, Table IV-3 highlights the necessity for the adaptions we described in the previous section.

Table IV-3: Most frequent sentimental words: LM and BPW

This table presents the most frequent positive, negative and uncertain words occurring in the quarterly and annual reports. For the English versions of the reports, the LM dictionary (Loughran & McDonald, 2011) was utilized to classify the words with respect to their sentimental categories. For the German versions of the reports, we utilized our own (BPW) dictionary. Note that the use of quarterly and annual reports obliges us to control for certain combinations of "PROFITS", "GAINS", and "LOSSES" as not doing so would result in an overestimation of the English speeches' shares of positive and negative words compared to the German counterparts. See Appendix IV-1 for complete list of all combinations we controlled for. Further, for the same reason, we counted the terms "IMPAIRMENT LOSS" and "IMPAIRMENT LOSSES" as one negative occurrence. The combined terms "IMPAIRMENT LOSSES" and "IMPAIRMENT LOSS" occurred 9,533 times and 2,776 times, respectively.

	English (LM)				German (BPW)			
Rank	Word	Total	%	cumulative %	Word	Total	%	cumulative %
Panel A: Most common negative words								
1	LOSS	37,522	7.6%	7.6%	VERLUSTE	20,624	4.2%	4.2%
2	LOSSES	21,378	4.3%	11.9%	GEGEN	19,371	3.9%	8.1%
3	AGAINST	19,898	4.0%	15.9%	BETRUG	17,987	3.7%	11.8%
4	NEGATIVE	17,198	3.5%	19.4%	RÜCKGANG	15,113	3.1%	14.9%
5	IMPAIRMENT	16,690	3.4%	22.8%	VERPFLICHTUNGEN	13,231	2.7%	17.6%
6	CLAIMS	14,852	3.0%	25.8%	VERLUST	11,191	2.3%	19.8%
7	DECLINE	12,907	2.6%	28.4%	WERTMINDERUNGEN	9,567	1.9%	21.8%
8	RESTRUCTURING	11,353	2.3%	30.7%	SCHADEN	9,245	1.9%	23.7%
9	DISCONTINUED	11,297	2.3%	33.0%	ERMITTLUNG	7,740	1.6%	25.2%
10	CRISIS	9,748	2.0%	35.0%	VERFÜGUNG	6,992	1.4%	26.6%
All		494,055				491,805		
Panel B: Most common positive words								
1	POSITIVE	25,526	6.7%	6.7%	POSITIVE	10,071	2.6%	2.6%
2	OPPORTUNITIES	20,339	5.3%	12.0%	ERTRAG	8,847	2.3%	4.8%
3	EFFECTIVE	14,109	3.7%	15.7%	POSITIV	8,650	2.2%	7.0%
4	ABLE	13,853	3.6%	19.3%	POSITIVEN	8,164	2.1%	9.1%
5	BENEFIT	13,245	3.5%	22.8%	ERREICHEN	6,749	1.7%	10.8%
6	IMPROVED	11,148	2.9%	25.7%	ERREICHT	6,529	1.7%	12.5%
7	GAINS	10,242	2.7%	28.4%	ERFOLG	6,191	1.6%	14.1%
8	ACHIEVED	10,222	2.7%	31.0%	VERBESSERUNG	6,102	1.6%	15.6%
9	SUCCESS	9,186	2.4%	33.5%	STEIGERUNG	5,911	1.5%	17.1%
10	LEADING	9,157	2.4%	35.8%	ERFOLGREICH	5,468	1.4%	18.5%
All		382,246				392,469		
Panel C: Most common uncertain words								
1	RISK	104,547	25.9%	25.9%	RISIKEN	45,179	11.8%	11.8%
2	RISKS	64,677	16.0%	41.9%	KANN	23,513	6.2%	18.0%
3	INTANGIBLE	26,516	6.6%	48.5%	RISIKO	19,699	5.2%	23.2%
4	COULD	15,034	3.7%	52.2%	IMMATERIELLE	18,788	4.9%	28.1%
5	POSSIBLE	14,195	3.5%	55.7%	ETWA	10,236	2.7%	30.8%
6	VARIABLE	11,941	3.0%	58.7%	ANNAHMEN	9,692	2.5%	33.3%
7	ASSUMPTIONS	11,032	2.7%	61.4%	PROGNOSE	6,318	1.7%	35.0%
8	APPROXIMATELY	10,192	2.5%	63.9%	MÖGLICH	6,287	1.6%	36.6%
9	EXPOSURE	8,512	2.1%	66.0%	IMMATERIELLEN	6,050	1.6%	38.2%
10	CONTINGENT	8,350	2.1%	68.1%	KÖNNTEN	5,996	1.6%	39.8%
All		403,844				381,551		

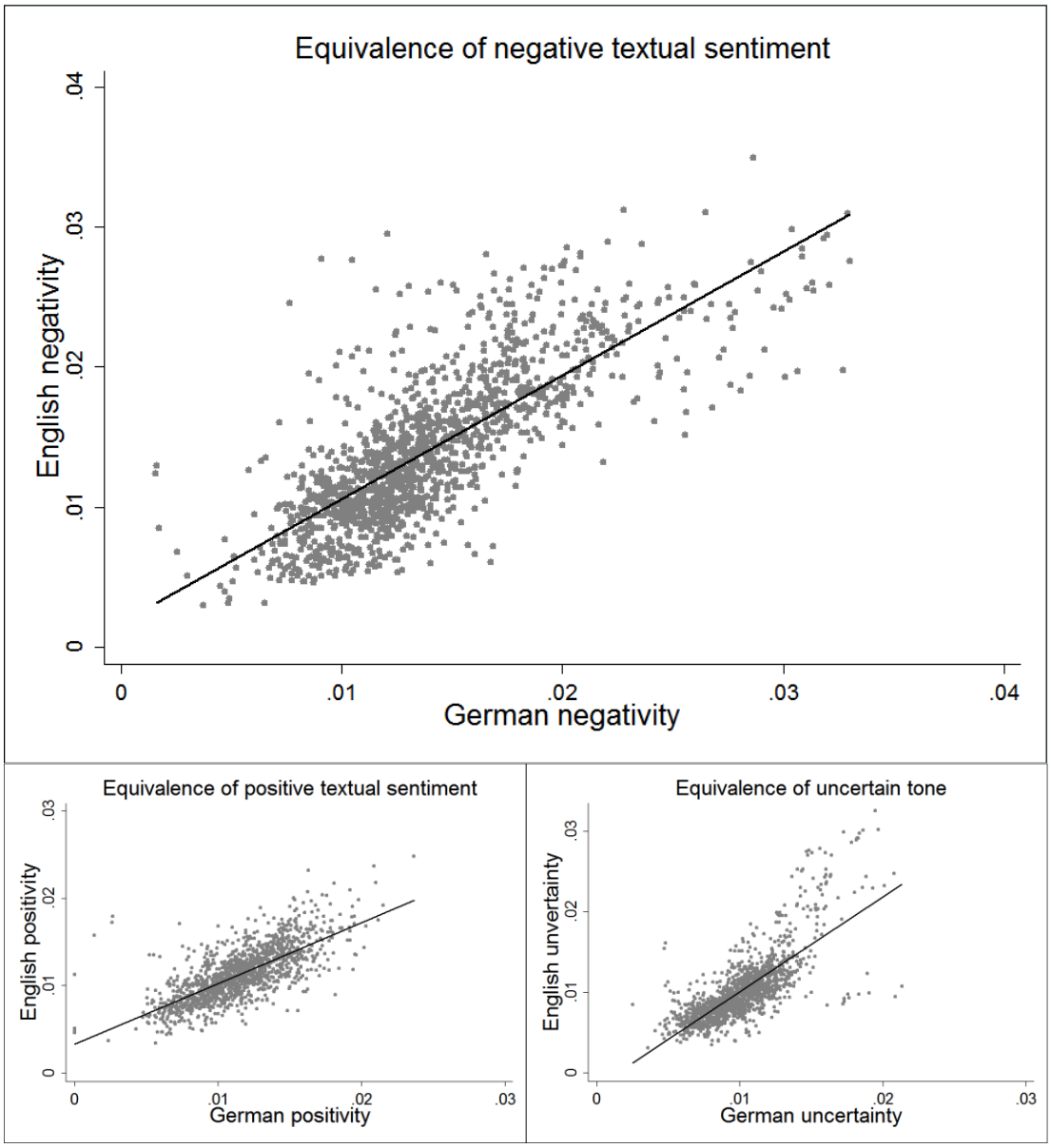
While the ten most frequent English negative, positive and uncertain words account for 35.0%, 35.8% and 68.1% of all negative, positive and uncertain words, their respective German counterparts only account for 26.6%, 18.5% and 39.8%, respectively. Apparently, as the total numbers of sentimental words are quite comparable between the sentimental categories, the German report's textual sentiment seems to be spread more widely among inflections. This result underlines that the sentiment in the German texts would have been strongly underestimated without the adjustments to the translation process and thus supports the necessity of the described adaptations in the creation of word lists for sentiment analyses.[10]

To further shed light on the equivalence of our BPW dictionary to the LM dictionary, we follow Wolf et al. (2008) and estimate simple pairwise correlations, Spearman rank correlations and intra-class correlations (ICC[3,2]) according to Shrout and Fleiss (1979), using our sample of 1,402 corresponding German and English quarterly and annual reports. Figure IV-1 graphically plots the 1,402 German and English reports' negative, positive and uncertain share of words and provides evidence of a strong and positive correlation. Table IV-4 displays summary statistics on the English and German reports' sentimental share of words. On average, the English versions of the reports, which are assessed with the LM dictionary, contain 1.41% negative, 1.13% positive and 1.02% uncertain words. These numbers are quite similar compared to the German reports which, using the BPW dictionary, contain 1.40% negative, 1.15% positive and 1.01% uncertain words.

[10] Loughran and McDonald (2011) give an account of the most frequently employed negative words in their sample of 10-K documents. The results for our sample of English quarterly and annual reports are quite comparable: All but one of the ten most common negative English words in our texts appear among the 30 most common negative words in Loughran and McDonald (2011). Furthermore, the distribution of the most common words within both samples appears to be quite comparable. While the ten most common negative words in our sample account for 35.0% of the negative word count, the ten most common negative words in Loughran and McDonald (2011) account for 33.8%.

Figure IV-1: Correlation plots of quarterly and annual reports

This figure plots the share of negative, positive and uncertain words for the quarterly and annual reports. The x-axis depicts the sentiment inherent the German version of the report as measured by our dictionary (BPW) and the y-axis depicts the sentiment inherent the English version of the report as measured by the dictionay by Loughran and McDonald (2011) (LM). The solid line reflects a linear regression.

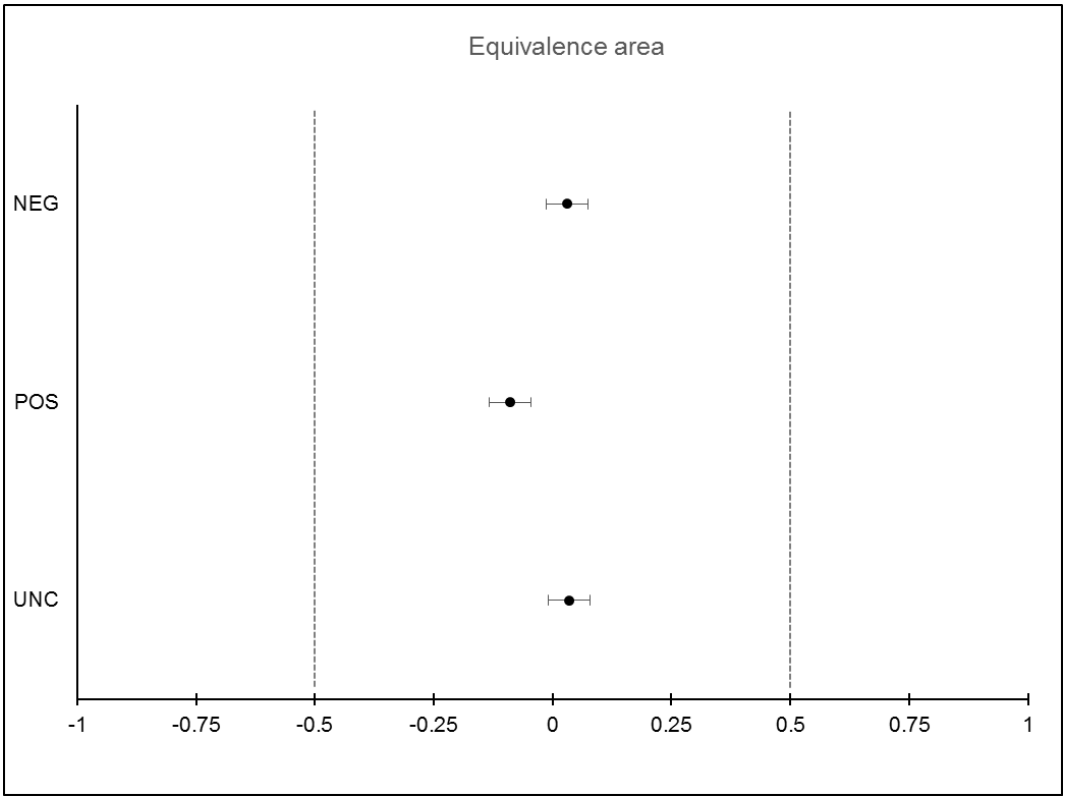


We assess the significance of these differences using two-sided equivalence testing following Blair and Cole (2002). As in Wolf et al. (2008), the area of insignificant deviation is set to 0.5 standard deviations around the zero difference between the standardized English original category and the German translation. In a two-sided approach, 90% confidence intervals are calculated for all category differences and compared in terms of their inclusion in the equivalence area. Fig-

ure IV-2 shows the results of the two-sided equivalence testing. All three categories are strikingly within the equivalence area, providing support for the equivalence of our BPW dictionary to the LM dictionary.

Figure IV-2: Equivalence tests after Blair and Cole (2002)

Figure II shows the results of the two-sided equivalence tests after Blair and Cole (2002). The area of insignificant deviation is it set to 0.5 standard deviations around the zero difference between the standardized English original category and the German translation. In a two-sided approach, 90% confidence intervals are calculated for all category differences, which corresponds to the calculation of two 95% confidence intervals, and compared in terms of their inclusion in the equivalence area.



Finally, Table IV-4 provides pairwise correlations, Spearman rank correlations and intraclass correlations (ICC[3,2]) between the English and German reports' shares of sentimental words. All correlations are considerably above 0.7, where values above 0.7 are considered to indicate reliability (Wolf et al., 2008).[11]

[11] We retest the BPW dictionary's equivalence to the LM dictionary not making any exemption from the word independency assumption explained in Appendix IV-1. The results are presented in Appendix IV-2 and support the dictionaries' equivalence.

Table IV-4: English vs. German textual sentiment: Reports

This table presents summary statistics for the quarterly and annual reports' shares of sentimental words with respect to the dictionary by Loughran and McDonald (2011) (LM) for the English versions of the reports and with respect to our (BPW) dictionary for the German versions of the reports. Further, this table shows simple pairwise correlations, Spearman rank correlations and intra-class correlations (ICC) after Shrout and Fleiss (1979) between the English and German textual sentiment with respect to the negative, positive and uncertainty wordlists by Loughran and McDonald (2011) and our adapted dictionary, respectively.

	English (LM)					German (BPW)							
	Mean [%]	Median [%]	SD [%]	Min [%]	Max [%]	Mean [%]	Median [%]	SD [%]	Min [%]	Max [%]	Pairwise Corr.	Spearman Corr.	ICC[3,2]
NEG	1.41	1.32	0.54	0.29	3.49	1.40	1.31	0.47	0.16	3.30	0.769	0.779	0.865
POS	1.13	1.10	0.30	0.34	2.48	1.15	1.14	0.31	0.00	2.37	0.725	0.734	0.840
UNC	1.02	0.96	0.40	0.31	3.25	1.01	1.01	0.26	0.25	2.13	0.752	0.774	0.811

4.2. Out of sample equivalence

In this section, we examine whether the BPW dictionary's equivalence holds with business-specific documents other than quarterly and annual reports. For this purpose, we re-estimate Table IV-3 and Table IV-4 using a sub-sample of the texts used in Bannier et al. (2017), who manually collect CEO speeches held at German DAX and MDAX companies' annual shareholder meetings from the companies' homepages. Our sub-sample consists of 270 speeches that are both available in English as well as in German. We conduct the same text processing steps as with the quarterly and annual reports, however, we do not apply any modifications to the bag-of-words model and its assumption of word independence. Summary statistics of the CEO speeches are provided in Table IV-5.

Table IV-5: Summary statistics of the CEO speeches

This table presents summary statistics of our sample of corresponding 270 English and German CEO speeches which were gathered from the companies' homepages.

			Words per document					Individual words per document				
	Reports	Total Words	Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
ENG	270	892,557	3,306	3,181	1,056	1,172	6,176	979	966	224	447	1,635
GER	270	931,213	3,449	3,343	1,047	1,327	6,392	1,126	1,114	265	530	1,835

Table IV-6 shows the CEO speeches' textual sentiment measured by the LM dictionary (English version) and by the BPW dictionary (German version). Unlike quarterly and annual reports (Table IV-4), CEO speeches appear to contain a considerably higher share of positive words. Specifically, while the quarterly and annual reports' English (German) versions share of positive words is 1.13% (1.15%), the respective CEO speeches' share is 3.14% (3.02%). The higher positivity does not come as a surprise, because CEOs have already been shown to

adjust their language in order to influence the audience. For example, Arslan-Ayaydin et al. (2015) find that CEOs inflate the use of positive language in earnings press releases, especially when their compensation is equity-based. Table IV-6 further shows the results for the simple pairwise correlations, Spearman rank correlations and ICC[3,2] using the 270 corresponding CEO speeches. The results are even stronger compared to the findings from the quarterly and annual reports, providing further support for our dictionary's equivalence to the LM dictionary.

Table IV-6: English vs. German textual sentiment: CEO speeches

This table presents summary statistics for our sample of 270 corresponding German and English CEO speeches' shares of sentimental words with respect to the dictionary by Loughran and McDonald (2011) for the English versions of the speeches and with respect to the BPW dictionary for the German versions. Further, this table shows simple pairwise correlations, Spearman rank correlations and intra-class correlations (ICC[3,2]) after Shrout and Fleiss (1979) between the English and German textual sentiment with respect to the negative, positive and uncertain wordlists dictionary by Loughran and McDonald (2011) and our adapted dictionary, respectively.

	English LM					German BPW							
	Mean [%]	Median [%]	SD [%]	Min [%]	Max [%]	Mean [%]	Median [%]	SD [%]	Min [%]	Max [%]	Pairwise Corr.	Spearman Corr.	ICC[3,2]
NEG	1.37	1.24	0.67	0.23	4.44	1.14	1.05	0.55	0.24	3.24	0.875	0.868	0.924
POS	3.14	3.22	0.71	1.55	5.29	3.02	2.99	0.64	1.66	5.42	0.841	0.845	0.911
UNC	0.72	0.67	0.31	0.08	2.08	0.77	0.74	0.31	0.14	1.82	0.749	0.735	0.856

4.3. The BPW dictionary vs. general language dictionaries

Table IV-7 compares the number of words in the BPW dictionary with the two general German dictionaries SENTIWS and LIWC. The SENTIWS dictionary comprises about 1.5 times more negative and about 6.9 times more positive words compared to our dictionary. Although all dictionaries appear quite comprehensive, the number of common words included in both dictionaries is quite low. The SENTIWS dictionary shares only 2,179 of the BPW's negative words and 993 of the BPW's positive words. Looking at the LIWC, the absolute number of included words is substantially lower compared to the BPW and the LIWC shares only 437 negative words and 247 positive words with the BPW. This is because the LIWC includes word stems, which account for several inflections of a word. As a consequence, the 'true' number of words in the LIWC as well as the 'true' match between the LIWC and the BPW is likely to be higher. To yield results comparable to the BPW and SENTIWS dictionaries, we therefore use a stemming algorithm by Caumanns (1999) on our sample of quarterly and annual reports before gauging their textual sentiment using the LIWC.

Table IV-7: Number of words in wordlists

This table shows the number of words in BPW, SENTIWS and LIWC dictionaries' positive and negative wordlists. Note that the LIWC, unlike the BPW and SENTIWS dictionaries, does not include broad sets of inflections but word stems which account for several inflections.

	BPW	SENTIWS		LIWC	
	No. of words	No. of words	Matching words	No. of words	Matching words
Negative	10,147	15,466	2,179	1,049	437
Positive	2,223	15,536	993	646	247

Table IV-8 shows the quarterly and annual reports' ten most frequent negative and positive words according to the SENTIWS and LIWC dictionaries. Panel A reveals that SENTIWS contains a similar number of negative words (451,933) compared to the BPW (491,805)[12]. However, SENTIWS includes a number of words that are inevitably used in business-related documents but are not necessarily negatively connoted. As examples, consider "ABSCHREIBUNGEN" (depreciation), "RISIKO" (risk), or "SCHULDEN" (debt). These three words alone account for 13.7% of negatively assessed words by SENTIWS. Of the ten most frequent negative words according to the LIWC, only two are not included in the BPW. However, this cannot be seen as an indication for a good match between the BPW and the LIWC, as the LIWC in total assesses only 142,710 words to be negative which is substantially lower compared to the BPW and SENTIWS dictionaries. As the BPW and LIWC assess quite comparable numbers of words as negative and the LIWC seems to miss a great share of negative words, our results rather indicate that the LIWC underestimates the negativity of business-related documents.

Panel B of Table IV-8 shows the quarterly and annual reports' ten most frequent positive words according to the SENTIWS and LIWC dictionaries. Compared to the BPW, it can be inferred that the general language dictionaries assess a considerable larger number of words as positive. More precisely, the SENTIWS (LIWC) yields 1,853,753 (461,341) positive words, while the BPW assesses 392,469 words as positive (Table IV-3). However, Panel B of Table IV-8 also shows that both general German dictionaries include words that are likely to be mis-specified in business-specific documents: "LEISTUNG(EN)" (service(s)), or "GEWINN" (profit), which represent the most frequent and second most frequent positive words according to SENTIWS and LIWC, are inevitably used within this context without necessarily having a positive connotation. Furthermore, the SENTIWS' ten most frequently employed positive words include "EIGENKAPITAL" (equity), "ANTEIL(E)" (share(s)), "INVESTITIONEN" (investments), "AK-

[12] See Table IV-3.

TIVITÄTEN (activities), and "WACHSTUM" (growth), which might not necessarily be considered as positive in business-specific documents. As a consequence, eight of the ten most frequent positive words in the SENTIWS dictionary seem to be mis-specified in the business context. These results indicate that both general language dictionaries, particularly the SENTIWS, might overestimate the positive sentiment of business-related text documents.

Table IV-8: Most frequent sentimental words: SENTIWS and LIWC

This table shows the quarterly and annual reports' most frequent negative and positive words with respect to the SENTIWS and LIWC dictionaries. Note that the LIWC contains word stems rather than comprehensive sets of inflections as the BPW and SENTIWS. Thus, we use a stemming algorithm by Caumanns (1999) on our sample of reports before gauging the textual sentiment using the LIWC.

Panel A: Negative textual sentiment

	SENTIWS				LIWC			
No.	Word	#	%	Cum. %	Word	#	%	Cum. %
1	ABSCHREIBUNGEN	26,526	5.87%	5.87%	✓VERLUST*	34,621	24.26%	24.26%
2	ENDE	23,638	5.23%	11.10%	✓BETRUG*	23,408	16.40%	40.66%
3	RISIKO	19,699	4.36%	15.46%	✓SCHULD*	18,425	12.91%	53.57%
4	✓BETRUG	17,987	3.98%	19.44%	✓SCHWACH*	7,183	5.03%	58.61%
5	SCHULDEN	15,684	3.47%	22.91%	AUFGAB*	6,306	4.42%	63.03%
6	✓RÜCKGANG	15,113	3.34%	26.25%	✓BELAST*	5,839	4.09%	67.12%
7	✓VERLUST	11,191	2.48%	28.73%	✓KLAG*	4,846	3.40%	70.51%
8	✓SCHADEN	9,245	2.05%	30.78%	✓SCHWIERIG*	4,081	2.86%	73.37%
9	TROTZ	8,332	1.84%	32.62%	FREMD*	3,653	2.56%	75.93%
10	✓NEGATIVE	6,564	1.45%	34.07%	✓NOT*	3,263	2.29%	78.22%
All		451,933				142,710		

Panel B: Positive textual sentiment

	SENTIWS				LIWC			
No.	Word	#	%	Cum. %	Word	#	%	Cum. %
1	LEISTUNGEN	35,333	1.91%	1.91%	GEWINN*	57,789	12.53%	12.53%
2	GEWINN	35,281	1.90%	3.81%	LEISTUNG*	42,921	9.30%	21.83%
3	ERTRÄGE	33,471	1.81%	5.61%	✓ERFOLG*	39,883	8.65%	30.47%
4	EIGENKAPITAL	30,854	1.66%	7.28%	✓POSITIV*	30,055	6.51%	36.99%
5	ANTEILE	30,772	1.66%	8.94%	BESTIMM*	23,635	5.12%	42.11%
6	INVESTITIONEN	25,618	1.38%	10.32%	✓ERREICH*	20,846	4.52%	46.63%
7	AKTIVITÄTEN	25,501	1.38%	11.70%	GUT*	18,839	4.08%	50.71%
8	NEUE	24,929	1.34%	13.04%	STARK*	18,633	4.04%	54.75%
9	ANTEIL	24,334	1.31%	14.35%	AKTIV*	15,947	3.46%	58.21%
10	WACHSTUM	23,917	1.29%	15.64%	WICHTIG*	15,593	3.38%	61.59%
All		1,853,753				461,341		

Table IV-9 shows Pearson and Spearman correlations between the reports' shares of sentimental words according to the BPW, SENTIWS and LIWC dictionaries. The correlation coefficients between the negative shares measured via

BPW and SENTIWS equal 0.749 (Pearson) and 0.740 (Spearman) and those between the respective positive shares equal 0.519 (Pearson) and 0.539 (Spearman). The correlation coefficients between the assessed textual sentiment of the BPW and SENTIWS are fairly high. Comparing the reports' shares of sentimental words according to the BPW and LIWC dictionaries, the correlation coefficients between the measures of negativity equal 0.519 (Pearson) and 0.539 (Spearman) and those between the measures of positivity equal 0.126 (Pearson) and 0.218 (Spearman). The correlation coefficients between the assessed textual sentiment of the BPW and LIWC dictionaries are lower compared to those between the BPW and SENTIWS dictionaries. Generally, our BPW dictionary seems to be more strongly correlated to the general language dictionaries when it comes to negative textual sentiment. This is in line with our previous finding, which indicates that the positive sentiment in business-specific text documents might be overestimated by the general language dictionaries.

Table IV-9: Correlations among sentiment measures

This table shows correlations among the sentiment measures. Note that the LIWC contains word stems rather than comprehensive sets of inflections as the BPW and SENTIWS. Thus, we use a stemming algorithm by Caumanns (1999) on our sample of reports before gauging the textual sentiment using the LIWC. Pearson correlations are below the diagonal, Spearman correlations are above the diagonal. P-values in parentheses.

	Negative			Positive		
	BPW	SENTIWS	LIWC	BPW	SENTIWS	LIWC
Negative						
BPW		0.740	0.562	-0.121	-0.078	-0.028
		(0.000)	(0.000)	(0.000)	(0.004)	(0.295)
SENTIWS	0.749		0.578	-0.126	-0.085	0.046
	(0.000)		(0.000)	(0.000)	(0.001)	(0.085)
LIWC	0.539	0.554		-0.173	-0.208	0.146
	(0.000)	(0.000)		(0.000)	(0.000)	(0.000)
Positive						
BPW	-0.129	-0.122	-0.208		0.539	0.218
	(0.000)	(0.000)	(0.000)		(0.000)	(0.000)
SENTIWS	-0.073	-0.038	-0.186	0.519		0.317
	(0.006)	(0.160)	(0.000)	(0.000)		(0.000)
LIWC	0.013	0.072	0.181	0.126	0.402	
	(0.623)	(0.007)	(0.000)	(0.000)	(0.000)	

5. Conclusion

We introduce a German adaptation of the widely-used dictionary by Loughran and McDonald (2011) that has been developed for analysis of business-specific text documents. Following the methodology of Wolf et al. (2008), we translate and adjust the positive, negative and uncertainty wordlists of the LM dictionary. We test the equivalence of our adaptation using a sample of 1,402 quarterly and

annual reports of German DAX and MDAX companies which are available in German and English. Our results provide broad evidence for the equivalence and reliability of our adaptation to its English original. We compare the BPW dictionary to two existing general German dictionaries, the SENTIWS and the LIWC dictionaries. With regard to the assessment of negative textual sentiment, the BPW and SENTIWS dictionaries appear to be highly and positively correlated. The LIWC dictionary, in contrast, seems to underestimate the negative textual sentiment. With respect to the assessment of positive textual sentiment, our results indicate that the SENTIWS and the LIWC dictionaries are likely to overestimate the positive textual sentiment, as they include a number of words, that cannot unequivocally be considered as positive in business specific text documents.

We contribute to the existing literature by providing a tool that allows to evaluate the qualitative nature of information disclosed in German text documents using the dictionary-based approach. Our BPW dictionary is not only specifically suited to examine business specific documents but also derived independently of a given text sample and hence applicable to different text types and formats. Furthermore, our approach describes a framework for future adaptations of English dictionaries into other languages and thus paves the way not only for further German content analyses, but also for studies in other languages and, thus, cultural backgrounds.

6. References

Ammann, M., & Schaub, N. (2016). Social interaction and investing: Evidence from an online social trading network. *Working Paper*.

Arslan-Ayaydin, Ö., Boudt, K., & Thewissen, J. (2016). Managers set the tone: Equity incentives and the tone of earnings press releases. *Journal of Banking & Finance*, 72, 132-147.

Bannier, C. E., Pauls, T., & Walter, A. (2017). CEO-Speeches and stock returns. *Working Paper*.

Blair, C., & Cole, S. R. (2002). Two-sided equivalence testing of the difference between two means. *Journal of Modern Applied Statistical Methods*, 1(1), 139-142.

Caton, S., Hall, M., & Weinhardt, C. (2015). How do politicians use Facebook? An applied social observatory. *Big Data & Society*.

Caumanns, J. (1999). A fast and simple stemming algorithm for German words. Freie Universität Berlin, Fachbereich Mathematik und Informatik Ser. B, Informatik: 99-16. Berlin: Freie Univ. Fachbereich Mathematik und Informatik.

Davis, A. K., Ge, W., Matsumoto, D., & Zhang, J. L. (2015). The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies*, 20(2), 639-673.

Davis, A. K., Piger, J. M., & Sedor, L. M. (2012). Beyond the numbers: Measuring the information content of earnings press release language. *Contemporary Accounting Research*, 29(3), 845-868.

Davis, A. K., & Tama-Sweet, I. (2012). Managers' use of language across alter-native disclosure outlets: Earnings press releases versus MD&A. *Contemporary Accounting Research*, 29(3), 804-837.

Doran, J. S., Peterson, D. R., & Price, S. M. (2012). Earnings conference call content and stock price: The case of REITs. *The Journal of Real Estate Finance and Economics*, 45(2), 402-434.

Ferris, S. P., Hao, Q., & Liao, M.-Y. (2013). The effect of issuer conservatism on IPO pricing and performance. *Review of Finance*, 17(3), 993-1027.

García, D. (2013). Sentiment during recessions. *The Journal of Finance*, 68(3), 1267-1300.

Haselmayer, M., & Jenny, M. (2016). Sentiment analysis of political communication. Combining a dictionary approach with crowdcoding. *Quality & Quantity*. doi:10.1007/s11135-016-0412-4.

Hawkins, J. A. (2015). A comparative typology of English and German: Unifying the contrasts (1st ed., Croom Helm, London, 1986). Routledge library editions : English language: Vol. 10. London [u.a.]: Routledge.

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *Journal of Business Communication*, 45(4), 363-407.

Henry, E., & Leone, A. J. (2016). Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. *The Accounting Review*, 91(1), 153-178.

Jacobi, C., Kleinen-von Königslöw, K., & Ruigrok, N. (2016). Political news in online and print newspapers. *Digital Journalism*, 4(6), 723-742.

Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3), 712-729.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33, 171-185.

König, E., & Gast, V. (2012). Understanding English-German contrasts (3rd Ed.). Grundlagen der Anglistik und Amerikanistik: Vol. 29. Berlin: Schmidt.

Larcker, D. F., & Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, 50(2), 495-540.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35-65.

Loughran, T., & McDonald, B. (2015). The use of word lists in textual analysis. *Journal of Behavioral Finance*, 16(1), 1-11.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187-1230.

Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. Cambridge, Mass.: MIT Press.

Mengelkamp, A., Hobert, S., & Schumann, M. (2015). Corporate credit risk analysis utilizing textual user generated content - A Twitter based feasibility study. *Working Paper*.

Mengelkamp, A., Schumann, M., & Wolf Sebastian. (2016). Data driven creation of sentiment dictionaries for corporate credit risk analysis. Proceedings of the 22. Americas Conference on Information Systems (AMCIS), 1-8.

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36(4), 992-1011.

Ramírez-Esparza, N., Pennebaker, J. W., García, F. A., & Suriá Martínez, R. (2007). La psicología del uso de las palabras: un programa de computadora que analiza textos en español. *Revista Mexicana de Psicología*, 24(1), 85-99.

Remus, R., Quasthoff, U., & Heyer, G. (2010). SentiWS - A publicly available German-language resource for sentiment analysis. LREC. 2010.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420-428.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis. MIT Press.

Tetlock, P. C. (2007). Giving content to investor sentiment. The role of media in the stock market. *The Journal of Finance*, 62(3), 1139-1168.

Tetlock, P. C., Saar-Tsechansky, M., & MacsKassy, S. (2008). More than words. Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3), 1437-1467.

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., & Kordy, H. (2008). Computergestützte quantitative Textanalyse. *Diagnostica*, 54(2), 85-98.

Zehe, A., Becker, M., Hettinger, L., Hotho, A., & Reger, I. (2016). Prediction of happy endings in German novels based on sentiment information. Proceedings of the Workshop on Interactions between Data Mining and Natural Language Processing 2016, 9-16.

Zijlstra, H., van Meerveid, T., & van Middendorp, H. (2004). De Nederlandse versie van de 'Linguistic Inquiry and Word Count' (LIWC). *Gedrag & Gezondheid*. (32), 271-281.

7. Appendix

Appendix IV-1: Adjustments to word independency assumption

This table lists word combinations we controlled for in our equivalence tests. While the bag-of-words model generally assumes word independence, the evaluation of quarterly and annual reports obliges us to control for certain combinations of words. A company's "GAINS AND LOSSES" or "PROFITS AND LOSSES" are frequently mentioned without negative or positive connotation in the quarterly reports. This would, in comparison to the German documents where the equivalent "GEWINN- UND VERLUSTRECHNUNG" is not included in the BPW dictionary, lead to a more extreme assessment of the English documents' positivity and negativity. Thus, we identify 40 combinations of the words "GAIN(S)" and "LOSSE(S)" as well as "PROFIT(S)" and "LOSSE(S)" and exclude them from the equivalence analyses. Likewise, the terms "IMPAIRMENT LOSS" and "IMPAIRMENT LOSSES" would account for two negative words while the German counterparts "WERTMINDERUNGSVERLUST" and "WERTMINDERUNGSVERLUSTE" would only account for one negative word. As this would also lead to an overestimation of the English documents' negativity compared to their German counterparts, we counted "IMPAIRMENT LOSS" and "IMPAIRMENT LOSSES" each as one negative word for our equivalence analyses. Note that we also controlled for different number of spaces between the combinations.

GAINS & LOSSES	PROFITS & LOSSES	LOSSES & GAINS	LOSSES & PROFITS
GAINS/LOSSES	PROFITS/LOSSES	LOSSES/GAINS	LOSSES/PROFITS
GAINS (LOSSES)	PROFITS (LOSSES)	LOSSES (GAINS)	LOSSES (PROFITS)
GAINS AND LOSSES	PROFITS AND LOSSES	LOSSES AND GAINS	LOSSES AND PROFITS
GAINS OR LOSSES	PROFITS OR LOSSES	LOSSES OR GAINS	LOSSES OR PROFITS
GAIN & LOSS	PROFIT & LOSS	LOSS & GAIN	LOSS & PROFIT
GAIN/LOSS	PROFIT/LOSS	LOSS/GAIN	LOSS/PROFIT
GAIN (LOSS)	PROFIT (LOSS)	LOSS (GAIN)	LOSS (PROFIT)
GAIN AND LOSS	PROFIT AND LOSS	LOSS AND GAIN	LOSS AND PROFIT
GAIN OR LOSS	PROFIT OR LOSS	LOSS OR GAIN	LOSS OR PROFIT

Appendix IV-2: English vs. German textual sentiment: No adjustment

This table presents summary statistics for the 1,402 quarterly and annual reports' shares of sentimental words with respect to the dictionary by Loughran and McDonald (2011) for the English versions of the reports and with respect to the BPW dictionary for the German versions of the reports. Further, this table shows simple pairwise correlations, Spearman rank correlations and intra-class correlations (ICC[3,2]) after Shrout and Fleiss (1979) between the English and German textual sentiment with respect to the negative, positive and uncertain wordlists dictionary by Loughran and McDonald (2011) and our adapted dictionary, respectively. For the analysis in this table, we do not make any exception from the word independence assumption.

	English LM					German BPW							
	Mean [%]	Median [%]	SD [%]	Min [%]	Max [%]	Mean [%]	Median [%]	SD [%]	Min [%]	Max [%]	Pairwise Corr.	Spearman Corr.	ICC[3,2]
NEG	1.58	1.47	0.57	0.34	3.84	1.40	1.31	0.47	0.16	3.30	0.779	0.779	0.863
POS	1.27	1.25	0.29	0.40	2.53	1.15	1.14	0.31	0.00	2.37	0.676	0.680	0.806

V. CEO-Speeches and stock returns

Co-authors: Christina E. Bannier, Andreas Walter
Own share: 70%

This paper was presented on the following refereed conferences/ workshops:

- 79th Annual Meeting of the German Academic Association for Business Research (VHB), St. Gallen, Switzerland, 2017.*
- International Conference on Discourse approaches to financial communication (DAFC), Lugano, Switzerland, 2017.*
- PhD Workshop of the German Finance Association (DGF), Bonn, Germany, 2016.

(* denotes forthcoming presentations)

This paper was presented on the following non-refereed conferences/ workshops:

- GGS doctoral seminar, Gießen, Germany, 2016.

CEO Speeches and Stock Returns

Christina E. Bannier[a] Thomas Pauls[b] Andreas Walter[c]

Abstract - We analyze the market reaction to the sentiment of the CEO speech at the Annual General Meeting (AGM). As the AGM is typically preceded by several information disclosures, the CEO speech may be expected to contribute only marginally to investors' decision-making. Surprisingly, however, we observe from the transcripts of 338 CEO speeches of German corporates between 2008 and 2016 that their sentiment is significantly related to abnormal stock returns and trading volumes following the AGM. Using a novel business-specific German dictionary based on Loughran and McDonald (2011), we find a negative association of the post-AGM returns with the speeches' negativity and a positive association with the speeches' relative positivity (i.e. positivity relative to negativity). Relative positivity moreover corresponds with a lower trading volume in a short time window surrounding the AGM. Investors hence seem to perceive the sentiment of CEO speeches at AGMs as a valuable indicator of future firm performance.

Keywords: Textual sentiment, CEO speeches, market efficiency, textual analysis, annual general meeting

JEL-Codes: G02, G12, G14

[a] Department of Banking & Finance, University of Gießen, Licher Str. 62, 35394 Gießen, Germany. Christina.Bannier@wirtschaft.uni-giessen.de.

[b] Department of Financial Services, University of Gießen, Licher Str. 74, 35394 Gießen, Germany. Thomas.Pauls@wirtschaft.uni-giessen.de.

[c] Department of Financial Services, University of Gießen, Licher Str. 74, 35394 Gießen, Germany. Andreas.Walter@wirtschaft.uni-giessen.de.

1. Introduction

Companies distribute information to relevant stakeholders by various means. Recent research has acknowledged the value not only of quantitative data disclosures but also of qualitative information, predominantly in the form of the textual sentiment of business communication. Sentiment is typically examined via content analyses which have been applied on several types of business communication such as annual reports (Feldman et al., 2008; Jegadeesh & Wu, 2013; Loughran & McDonald, 2011, 2015), earnings press releases (Davis et al., 2012; Davis & Tama-Sweet, 2012; Henry, 2008; Henry & Leone, 2016; Huang et al., 2014), IPO prospectuses (Demers & Vega, 2008; Ferris et al., 2013; Jegadeesh & Wu, 2013), CEO letters (Boudt & Thewissen, 2016), and earnings conference calls (Davis et al., 2015; Doran et al., 2012; Larcker & Zakolyukina, 2012; Price et al., 2012). In general, these studies find that qualitative information is indeed processed by investors and helps to predict future accounting returns, stock returns, stock volatility, and stock trading volume.[1]

Surprisingly, the Annual General Meeting (AGM) received only little attention so far and the CEO's speech held at the AGM hardly any. Only few studies investigate the market reaction to the AGM at all and those that do report inconclusive and partly diverging results. Firth (1981), for example, does not find a market reaction in terms of abnormal returns and trading volume. Brickley (1986) and Rippington and Taffler (1995) report only small price reactions around the AGM for US and UK firms, respectively. Olibe (2002) presents evidence of a minimal trading-volume response to UK companies' AGM. Martinez-Blasco et al. (2015) find no significant market reactions in Japan and Spain and only trading volume increases for US, UK, and French stocks. For German stocks, in contrast, they observe significant market responses to the AGM in terms of increased returns, return volatility and trading volume.

The generally weak market reaction to the AGM may be explained by the fact that the AGM is typically preceded by several information disclosures such as preliminary earnings announcements and the full release of the annual report. As a consequence, the AGM can hardly deliver any new quantitative information. However, to the best of our knowledge, no attempts have been made so far to investigate the qualitative content of the AGM and of the CEO's speech in particular. This is despite the fact that the AGM offers managers the rare opportunity to personally address the company's stockholders in order to share their views on the firm's prospects (Martinez-Blasco et al., 2015).

[1] See Kearney and Liu (2014) or Loughran and McDonald (2016) for a comprehensive overview.

The lack of studies on the qualitative content of CEO speeches is particularly surprising, since CEO communication in general has been shown to exhibit valuable qualitative information. For example, Arslan-Ayaydin et al. (2015) find that incentivized managers use positive words more aggressively in an attempt to influence share prices. Similarly, Boudt and Thewissen (2016) report that CEOs strategically present negative and positive words in CEO letters in order to prompt a more positive perception by the reader. Price et al. (2012) and Doran et al. (2012) show that the tone of earnings conference calls - which are typically conducted by the firm's top management team - is a significant predictor of subsequent returns and trading volume. We therefore hypothesize that CEO speeches held at AGMs contain valuable qualitative information that should influence the market reaction to the AGM. As Demers and Vega (2008) find that financial markets tend to incorporate qualitative information with delay, we furthermore presume investors to initially underreact to the speeches' sentiment so that the full market reaction will present itself only in a protracted time period after the AGM.

We test our hypothesis on the CEO speeches of publicly listed companies in Germany. We choose German firms as they regularly release the speeches' transcripts on their websites immediately after the AGM. US companies, in contrast, only rarely provide respective transcripts: While 72.50% of the German DAX and MDAX[2] companies offer transcripts, only 5.8% of all S&P 500 firms do so, rendering a meaningful empirical analysis on US data all but impossible.[3] We consider 338 CEO speeches of DAX and MDAX-listed corporations in Germany from 2008 to 2016. In a first step, we analyze whether AGMs systematically reveal new information per se and measure the financial market reaction subsequent to the AGM. Our univariate results show that AGMs do not seem to be followed by abnormal returns and we find a higher trading volume only in a short time window around the AGM.

In a second step, we examine whether the CEO speeches' sentiment at the AGM contains value-relevant information that is picked up by financial market participants. Sentiment is typically measured via a dictionary-based approach by assigning the words in a text or speech to different sentiment categories in accordance with a predefined dictionary (Manning & Schütze, 1999). Using a novel dictionary by Bannier et al. (2017), we gauge the sentiment of the CEO speeches

[2] The DAX and MDAX indices comprise the 80 largest German stock-listed companies in terms of order book volume and market capitalization. For more information on the indices, see http://www.dax-indices.com/EN/.

[3] Altogether, we were able to download only 54 speeches of US companies listed in the S&P 500.

and assess the financial market reaction to the AGM with respect to this sentiment. Our results show the sentiment to be significantly related to cumulative abnormal returns and trading volume. More precisely, we find the cumulative abnormal returns to decrease along with a speech's negativity and to increase with a speech's relative positivity, i.e., its positivity relative to negativity. Investigating the time structure of the sentiment's effect, we find that only a small part of the full market reaction occurs in the immediate vicinity around the AGM. Most of the market reaction, however, is observed in the time period between 2 and 30 days post AGM. This observation may be interpreted as an initial underreaction to the speeches' sentiment and could be an indication that qualitative information indeed needs more time to become fully incorporated in stock prices. Interestingly, the speeches' relative positivity is also significantly associated with a lower cumulative abnormal trading volume, but only in a short time window around the AGM. In summary, we find that a more positive relative to negative sentiment of a CEO speech goes along with higher cumulative abnormal returns and lower short-term cumulative abnormal trading volumes of the company's stock, whereas a lower positive relative to negative sentiment triggers lower returns and higher trading volumes.

Our paper's contribution to the existing literature is twofold. To begin with, we are the first to measure the sentiment of corporate texts in the German language using the business-specific dictionary introduced by Bannier et al. (2017). While the studies by Ammann and Schaub (2016) or Mengelkamp et al. (2015) also investigate the sentiment in German corporate texts, they either utilize only general German language dictionaries or ad-hoc dictionaries restricted to the respective sample of text documents at hand. The dictionary of Bannier et al. (2017), in contrast, is designed to capture the business-specific sentiment of any sample of German documents in a comprehensive way and follows the setup of the Loughran and McDonald (2011) dictionary for English documents.

As we are the first to employ this context-specific dictionary, we compare our results to those derived from using two general German language dictionaries. These are the "SentimentWortschatz" by Remus et al. (2010) and the German adaptation of the Linguistic Inquiry Word Count by Wolf et al. (2008). In line with content analyses on English documents (Henry & Leone, 2016; Loughran & McDonald, 2011, 2015; Price et al., 2012), we find the context-specific dictionary to be better suited for assessing the textual sentiment of business-related documents than general language dictionaries. Given the economic importance of firms in Germany and other German-speaking countries and the robust performance of the dictionary introduced by Bannier et al. (2017), the dictionary can hence be seen as a helpful tool to assess the qualitative information contained in

these firms' communication. We also check the robustness of our results to different word weighting schemes, i.e., equal weighting vs. inverse document frequency weighting as proposed by Loughran and McDonald (2011). We find no improvement from using inverse document frequency weighting, similar to Henry and Leone (2016). Finally, we determine which measure of textual sentiment is most appropriate to gauge the qualitative information within German text documents. Consistent with Price et al. (2012) and Henry and Leone (2016) for English content analyses, we find the measure of relative positivity, which combines both positive and negative sentiment, to perform better than the positivity or negativity measure in isolation.

The second and main contribution of our study, however, is to show that there is valuable qualitative information hidden in the annual get-together of managers and shareholders. Our results suggest that financial market participants do indeed pick up the qualitative information contained in the CEO's speech for their investment decisions. However, both negativity and relative positivity - as the two most meaningful sentiment categories - are incorporated in the stock price only with a certain delay: While in the short time period around the AGM the association between the speeches' relative positivity and cumulative abnormal returns is only weak, the major part of the market reaction occurs in the time period between 2 and 30 days after the AGM. At the same time, however, we find a significant association between the relative positivity and the cumulative abnormal trading volume solely in the short time window immediately surrounding the AGM. The comparably long-lasting impact of the CEO speeches' sentiment on stock returns hence seems to be accompanied by an attention-capturing effect on the trading volume that is, however, quickly evaporating.

The remainder of the paper is structured as follows. The next section reviews the literature on the information provided in AGMs as well as on content analyses. Further, it introduces the dictionary developed by Bannier et al. (2017). Section 3 describes our data and the methodology employed. Section 4 presents the respective results. Finally, Section 5 concludes.

2. Literature

2.1. Informational content of the annual general meeting

Companies typically release their annual results in three stages. First, a preliminary announcement is made including information about the company's profits, earnings per share, dividend per share, and sales turnover. A few weeks later, the company releases its annual report and finally, some weeks after that, the

company's AGM takes place. Accordingly, Firth (1981) finds that the preliminary announcement and the release of the annual report induce significant abnormal returns and trading volume, while he finds no such market reaction following the AGM. Hence, he concludes that the AGM does not seem to provide new information to financial markets. This is supported by García-Blandón et al. (2012) who evaluate the AGM's information value in Spain and find no market reaction at all. Brickley (1986), Rippington and Taffler (1995) and Olibe (2002) observe only small price and trade volume reactions around the AGM. The most comprehensive study on the AGM's information value has been conducted by Martinez-Blasco et al. (2015) on a sample of common- and civil-law countries. The authors examine changes in abnormal returns, return volatility, and trading volume. Their analysis reveals no market reaction in Japan and Spain, and only small increases in trading volume in the US, the UK, and in France. In Germany, in contrast, the authors observe significant increases in abnormal returns, return volatility, and trading volume following the AGM, indicating that the AGMs of German companies exhibit substantial new information.

Despite the mixed results, none of the earlier studies - to the best of our knowledge - attempts to investigate the source or type of any potential information disclosure at the AGM. This is surprising since the AGM is a rare opportunity for a firm's management to get into direct contact with its shareholders (Martinez-Blasco et al., 2015) and since there is plenty of evidence on qualitative information inherent in the language of CEOs. Arslan-Ayaydin et al. (2015) find that managers adjust their language to specific situations at hand and inflate the use of positive language the higher their fraction of equity-based compensation. Doran et al. (2012) and Price et al. (2012) report that conference calls' positive sentiment is a significant predictor of subsequent returns and trading volume. Mayew and Venkatachalam (2012) and Hobson et al. (2012) analyze conference call audio files using vocal emotion analysis software. They come to the conclusion that positive and negative emotions expressed in the voice of managers can be informative about the firm's financial future and potential financial misreporting. It is hence reasonable to believe that qualitative information may be contained in the AGM even though substantial quantitative information has already been distributed to investors prior to the meeting. We therefore assess whether this qualitative information is inherent in the verbal communication of the CEO at the AGM.

2.2. Dictionary-based approach

The dictionary-based approach has become a commonly used tool to measure the textual sentiment of various kinds of documents such as financial disclosures,

analyst reports, earnings press releases, IPO prospectuses, internet board postings, or newspaper articles (Kearney & Liu, 2014). The individual dictionaries typically include various wordlists with respect to sentimental categories such as negativity or positivity. Text documents with a comparably high share of, for example, negative words are then considered to be more pessimistic compared to text documents with a comparably high share of positive words (Loughran & McDonald, 2015).

Early content analyses of financial texts (Davis et al., 2012; Davis & Tama-Sweet, 2012; Feldman et al., 2008; Ferris et al., 2013; Henry & Leone, 2016; Kothari et al., 2009; Larcker & Zakolyukina, 2012; Tetlock, 2007; Tetlock et al., 2008) utilized general English dictionaries such as the Harvard University's General Inquirer IV-4[4] dictionary, the dictionaries included in the Diction[5] software, or the Linguistic Inquiry Word Count[6] software. Henry (2008) is the first to compose a dictionary explicitly designed to examine the tone of financial documents. Despite the comparably small number of words in her positive and negative word lists, various studies comment on the superiority of the dictionary presented by Henry (2008) over the Diction and General Inquirer dictionaries (Doran et al., 2012; Henry & Leone, 2016; Price et al., 2012). Based on this finding, Loughran and McDonald (2011) create a more comprehensive dictionary (hereafter LM dictionary) by evaluating all words that appear in at least 5% of the entire 10-K disclosure universe. The LM dictionary contains 2,329 negative and 354 positive words. To assess the quality of their dictionary, the authors show that 73.8% of the General Inquirer dictionary's negative words do not have a negative meaning in financial documents and, in later work, demonstrate that none of the most frequently occurring negative words in the 10-K disclosures are included in the Henry (2008) dictionary (Loughran & McDonald, 2015). Due to its comprehensiveness and its appropriateness for financial documents, the LM dictionary has become the most widely used dictionary in business research and has been used to assess the textual sentiment of 10-K filings (Loughran & McDonald, 2011), earnings conference calls (Davis et al., 2015), news articles (García, 2013), or IPO prospectuses (Ferris et al., 2013; Jegadeesh & Wu, 2013).[7]

[4] See http://www.wjh.harvard.edu/~inquirer/.

[5] See http://www.dictionsoftware.com/.

[6] See http://www.liwc.net.

[7] For a comprehensive overview of dictionaries used in content analyses, see Kearney and Liu (2014) and Loughran and McDonald (2016).

2.3. German language dictionaries

When it comes to the analysis of German text documents, two comprehensive general German language dictionaries but no business-specific dictionary exist: Remus et al. (2010) created the "SentimentWortschatz" (hereafter SENTIWS) dictionary, which is based on and extends the General Inquirer lexicon by Stone et al. (1966). SENTIWS has been used in studies of political communication (Haselmayer & Jenny, 2016), or art and literature (Zehe et al., 2016). The second general language dictionary was created by Wolf et al. (2008), who adapted the English version of the Linguistic Inquiry Word Count to the German language. Their dictionary (hereafter LIWC) puts special emphasis on analyzing essays in the context of expressive writing experiments, but has also been used in other research domains such as, for example, political analyses (Caton et al., 2015; Jacobi et al., 2016). However, with respect to business-related documents, there is no context-specific dictionary.

As many text documents containing relevant information on German companies are published exclusively in German, the absence of a context-specific dictionary in German is associated with very little research on German qualitative information. Rare exceptions are Ammann and Schaub (2016) and Mengelkamp et al. (2015), who investigate German corporate texts for their textual sentiment and utilize ad-hoc dictionaries that are constructed from - and thus restricted to - a given set of sample text documents. Similar to the studies conducted on English text documents, the authors also find that their ad-hoc dictionaries achieve more reliable results than the general German language dictionaries SENTIWS and LIWC.

In order to analyze German business-related texts comprehensively, Bannier et al. (2017) adapt the English business-specific dictionary by Loughran and McDonald (2011) to the German language. They follow the methodology of Wolf et al. (2008) and control for several linguistic issues such as inflections, compound words, or lexical morphology that are specific to the German language (König & Gast, 2012; Hawkins, 2015).[8] For a detailed explanation of the setup of word lists to measure sentiment in German corporate texts, see Bannier et al. (2017). The

[8] German speakers are forced to make certain inflectional distinctions which can regularly be left unspecified in English. Looking at verbs, for example, the German language distinguishes indicative and subjunctive forms whereas English employs a single form for both. Further, German verbs differ with respect to person and number, whereas the bare stem in English is used for all except the third person singular. As German nouns and adjectives need more inflections as well, a simple word-by-word translation of the LM dictionary will not fully cover the German inflectional morphology with the consequence of an underestimation of the German texts' sentiment.

authors also test the equivalence of their adaptation (hereafter BPW dictionary) using a broad sample of quarterly and annual reports of German companies that are available in German and English language.[9] The results show that all sentiment categories display high correlation and equivalence to their English counterparts, indicating the reliability of their adaptation.[10]

Table V-1 presents a brief comparison of the LM dictionary, the BPW dictionary and the two general German dictionaries, SENTIWS and LIWC, that allows to put the specificities of the German language into perspective and helps to see the differences between general and context-specific wordlists.

Table V-1: Dictionaries for content analysis

This table shows the number of words contained in the positive and negative wordlists of existing English and German language dictionaries for content analysis. Note that the LIWC contains word stems rather than comprehensive sets of inflections as LM, BPW, and SENTIWS.

	English	German		
	LM	BPW	SENTIWS	LIWC
Negative	2,354	10,147	15,466	1,049
Positive	354	2,223	15,536	646

Table V-1 shows that the German dictionaries' word lists contain far more individual words than the English LM dictionary. This is mainly due to the linguistic issues referred to above. However, even within the German language, there are strong differences between the dictionaries. Comparing the BPW to the SENTIWS dictionary reveals that SENTIWS includes about 50% more negative and about 700% more positive words than the BPW. Overall, SENTIWS contains as many negative as positive words. This stands in contrast to the other dictionaries, most obviously the LM dictionary, which contains a much smaller number of positive than negative words. Note that a direct comparison of the number of individual words between the BPW and the LIWC dictionaries is not feasible as LIWC includes word stems rather than inflections. However, both general German dictionaries are likely to include words that may misclassify sentiment in a business context. For example, "LEISTUNG(EN)" (service(s)), or "GEWINN" (profit), which are both classified as positive words by SENTIWS and LIWC, are regularly used in business documents without a necessarily positive connotation. Other examples such as "EIGENKAPITAL" (equity), "ANTEIL(E)" (share(s)),

[9] We estimate simple pairwise correlations, Spearman rank correlations, intra-class correlations Shrout and Fleiss (1979), and test the dictionaries' equivalence via two-sided equivalence testing following Blair and Cole (2002).

[10] For more information on the adaptation process and equivalence tests of the BPW dictionary, see Bannier et al. (2017).

"INVESTITIONEN" (investments), "AKTIVITÄTEN (activities), and "WACHSTUM" (growth) are also counted as generally positive, while this may not be the case in business-related documents. As a consequence, both general language dictionaries and particularly the SENTIWS word lists might overestimate the positive sentiment of business-related text documents. While the higher fit of context-specific dictionaries has already been confirmed by English language studies (Price et al., 2012; Loughran & McDonald, 2011, 2015; Henry & Leone, 2016), this issue is still unresolved in the German language. In the following analysis, we will therefore put some emphasis on evaluating the efficacy of the BPW dictionary relative to the two general language dictionaries, SENTIWS and LIWC, when employing the different dictionaries on the CEO speeches.

3. Data and methodology

3.1. Data and variable measurement

We attempt to capture the sentiment in CEO speeches held at German companies' AGMs and to assess whether this sentiment is associated with significant market reactions subsequent to the AGM. For that purpose, we gather the CEO speeches held at German DAX and MDAX companies' annual shareholder meetings from 2008 to 2016 by manually collecting transcripts from the companies' internet webpages. Our initial sample consists of 356 CEO speeches by 58 companies. We evaluate further documents, such as company charters, shareholder meeting invitations, and audio or video material from the companies' webpages, in order to confirm that the CEO speeches are indeed initially held in German. Based on this additional analysis, we exclude 18 speeches resulting in a final sample of 338 speeches.

Before we can segment the reports into vectors of word counts, we have to convert the documents, which are typically available in PDF file format, to TXT format. In this process, we also replace typographic ligatures and employ UTF-8 character encoding on all files in order to allow for German-specific characters such as 'Ä','Ü','Ö', or 'ß'. All characters are transformed into lower case and tokenized afterwards, whereby we define a token as any subsequent order of at least three alphabetic characters. In order to exclude potential spelling errors, we exclude tokens that do not occur in at least one percent of the speeches. After that, we apply a stop-word list on the reports to filter out words that might have important semantic functions, but rarely contribute information (Manning & Schütze, 1999). We use the stop-word list provided by Bannier et al. (2017) which includes common names, dates, numbers, geographic locations, currencies,

the names of German DAX and MDAX companies, popular German pre- and surnames, and the names of the largest German and European cities. Hereafter, the documents are transformed to word count vectors using the Rapidminer software.[11] In a final step, the CEO speeches' numbers of negative and positive words are counted with respect to the word lists of the BPW, SENTIWS and LIWC dictionaries.

Several measures to gauge textual sentiment have been utilized in the literature. Jegadeesh and Wu (2013) and García (2013) employ direct measures of positivity and find statistically significant market reactions. Loughran and McDonald (2011, 2016), however, point out that positive words are frequently used to frame negative words, whereas negative words are unambiguous in their usage. Consequently, Tetlock (2007) and Loughran and McDonald (2011) find little incremental information using only a positive wordlist and suggest using a documents' share of negative words to assess its textual sentiment. We therefore estimate the CEO speeches' share of negative words as follows:[12]

$$NEG\_BPW_j = \frac{NEGATIVE_j}{COUNT_j} * 100 \quad (1)$$

Here, $COUNT_j$ is the total number of words of CEO speech j and NEGATIVEj represents the number of negative words in CEO speech j with respect to the negative wordlist of the BPW dictionary. $NEG\_SENTIWS_j$ and NEG_LIWC_j are calculated analogously.

Recent studies point out, however, that recipients of financial documents might not consider positive and negative textual sentiment separately but rather in relation to each other. We therefore follow Henry (2008), Price et al. (2012), and Henry and Leone (2016) and estimate the CEO speeches' relative positivity (TONE) in the following way:

$$TONE\_BPW_j = \frac{POSITIVE_j - NEGATIVE_j}{POSITIVE_j + NEGATIVE_j} \quad (2)$$

Here, $POSITIVE_j$ is the number of positive words in CEO speech j with respect to the positive wordlist of the BPW dictionary. $TONE\_SENTIWS_j$ and

[11] The transformation to lower-case characters, the tokenization, the stop-word filtering and the generation of the word count vectors were conducted with the Rapidminer software. For more information, please see https://rapidminer.com/.

[12] We re-estimate our main-analysis grasping the CEO speeches sentiment using a measure of positivity. The results are shown in Table V-6.

$TONE\_LIWC_j$ are calculated analogously. The relative positivity measure - also referred to as tonality - hence combines the information of the negative and positive sentiment as it measures the positivity of speech j relative to its negativity. The TONE measures are scaled between -1 and 1, so that a purely positive CEO speech displays a score of 1, a purely negative speech a score of -1, and a neutral speech scores a 0.

In order to measure the stock price reaction subsequent to a CEO speech, we calculate Cumulative Abnormal Returns (CARs). For this, daily abnormal returns are calculated using the return of the CDAX[13] index as the expected return, which reflects the performance of the entire German equity market:

$$AR_{j,t} = R_{j,t} - R_{CDAX,t} \tag{3}$$

Here, $AR_{j,t}$ is the abnormal return on company j's stock at day t and $R_{j,t}$ is the actual return of company j's stock at day t. $R_{CDAX,t}$ is the return of the CDAX on day t. As Demers and Vega (2008) find that qualitative information is more difficult for market participants to process than quantitative information, we may expect any market reaction to the sentiment in CEO speeches to not be overly quick. We therefore examine the market reaction by cumulating the abnormal returns for each stock over a relatively long time period from day -1 to day 30, where 0 represents the day of the AGM at which speech j is held. To analyze the time structure of a potential market reaction in more detail, we then segregate this total time window into the three-day period around the AGM (-1,1) and the remaining period after the AGM (2,30). This approach should allow us to see whether the market reaction to the sentiment in CEO speeches operates in an immediate or a delayed fashion. We hence employ three CAR measures, estimated in the following way:

$$CAR(-1,30)_j = \sum_{t=-1}^{30} AR_{j,t} \tag{4}$$

$$CAR(-1,1)_j = \sum_{t=-1}^{1} AR_{j,t} \tag{5}$$

[13] The CDAX comprises the price development of all 852 German stocks across the Deutsche Börse's prime and general standard. For more information on the CDAX, see http://www.dax-indices.com/EN/.

$$CAR(2,30)_j = \sum_{t=2}^{30} AR_{j,t} \quad (6)$$

In addition to analyzing the CEO speeches' sentiment effect on stock prices, we also measure the effect on actual trading. For this purpose, we estimate the Cumulative Abnormal Trading Volume (CAV) following Barber and Odean (2008) and Price et al. (2012), where the Abnormal Trading Volume (AV) is in a first step calculated as follows:

$$AV_{j,t} = \frac{VOLUME_{j,t}}{\overline{VOLUME}_{j,t}} - 1 \quad (7)$$

Here, $VOLUME_{j,t}$ is the trading volume for company j at day t, and $\overline{VOLUME}_{j,t}$ is the mean trading volume for company j from t-252 to t-1. Consequently, a value of zero for the abnormal trading volume $AV_{j,t}$ indicates that a company's stock j was not traded abnormally at day t compared to the previous 252 days, i.e., over the last year. A positive value indicates that the stock was traded more than usual and a negative value indicates that the stock was traded less than usual. Analogously to abnormal returns, $AV_{j,t}$ is accumulated over day -1 to 30, CAV(-1,30), day -1 to 1, CAV(-1,1), and day 2 to 30, CAV(2,30).

3.2. Empirical approach

In a first univariate analysis, we sort the CEO speeches into quartiles with respect to the measures of textual sentiment and compare the mean and median CAR(-1,30), CAR(-1,1), and CAR(2,30) differences between the highest and lowest quartiles of textual sentiment. We then test the mean and median differences for statistical significance using t-tests and Wilcoxon rank sum tests, respectively. To check whether the univariate results of our sentiment measures hold in a multivariate setting, we then conduct cross-sectional OLS regressions with a comprehensive set of control variables of the following form:

$$CAR_j = \alpha_0 + \alpha_{1,i} * SENTIMENT_{i,j} + \alpha_{2,k} * CONTROLS_{k,j} + \varepsilon_j \quad (8)$$

Here, CAR_j is the measure of cumulative abnormal returns for CEO speech j, $SENTIMENT_{i,j}$ is a vector of the different sentiment measures i for speech j which are calculated as described above. $CONTROLS_{k,j}$ represents a vector of control variables for speech j which include the speech's length (COUNT), the speech's

share of individual words (IND), the earnings surprise (EPS_SURP), the dividend surprise (DIV_SURP_POS and DIV_SURP_NEG), the market capitalization (SIZE), market to book ratio (M2B), leverage (LEVERAGE), return on assets (ROA), return volatility (VOLATILITY), and trading volume (VOLUME).[14]

COUNT represents the CEO speeches' length in terms of the total number of words. IND is the number of individual words in a CEO speech divided by the speech's total number of words. The earnings surprise (EPS_SURP) of CEO speech j is estimated in accordance with Price et al. (2012) as the difference between the last reported earnings per share for the company at time t minus the latest reported earnings per share in the year prior to date t, divided by the stock price one year before t:

$$EPS\_SURP_j = \frac{EPS_j - EPS_{j,t-1}}{STOCKPRICE_{j,t-1}} * 100 \quad (10)$$

Here EPS_j is the most recent earnings-per-share release for the company at the time of speech j, $EPS_{j,t-1}$ is the most recent earnings-per-share release for the company one year before the day of speech j and $STOCKPRICE_{j,t-1}$ is the stock price of the company one year before the date of speech j. While the earnings surprise has been shown to affect returns and volatility following earnings announcements and earnings conference calls, we hypothesize that EPS_SURP should only have a limited effect on the CARs following the CEO speeches since the surprise is already known from the quarterly report und, thus, should already be incorporated in the stock price at the time the speech is held. We include the indicator variables DIV_SURP_POS and DIV_SURP_NEG to control for dividend surprises. Here, DIV_SURP_POS is equal to one if a company's dividend per share is increased compared to the previous year, zero otherwise, and DIV_SURP_NEG is equal to one if a company's dividend per share is decreased compared to the previous year, zero otherwise. In contrast to the earnings surprise, the dividend surprise might strongly influence the post AGM returns and trading volume, as the dividend is actually agreed on at the AGM. SIZE measures the company's equity market value at the day of the speech as the share price multiplied by the number of ordinary shares outstanding. It is displayed in Euro millions. We include the market to book ratio (M2B) to control for the company's growth opportunities. M2B is defined as the market value of the ordinary equity

[14] Note that we include COUNT, IND, SIZE and VOLUME in logarithmic format in the regressions.

divided by the balance sheet value of the ordinary equity in the company. We include ROA, LEVERAGE and VOLATILITY to control for a potentially higher information demand by investors which might result from low profitability, financial distress or other forms of uncertainty, respectively. ROA is estimated as net income divided by total assets times one hundred. LEVERAGE is calculated as total liabilities divided by total assets and VOLATILITY is estimated as the daily returns' standard deviation in the time window of minus 90 days to minus 10 days prior to the AGM. Finally, VOLUME describes the number of shares traded of a stock on the day of the shareholder meeting and is expressed in thousands. While our sentiment measures, COUNT, and IND are collected directly from the CEO speeches, the data to estimate the remaining control variables are gathered from Thompson Reuters Datastream. We repeat all previously described analyses, substituting CAV for CAR. In the multivariate analyses we then utilize the same set of control variables except for VOLUME.

3.3. Weighting scheme

The majority of studies employing the dictionary-based approach use equal weighting of individual words. This method values each individual word in a document equally and implies that a more frequent occurrence of a word indicates a higher importance.[15] However, as the impact of words might be diluted the more often they are used, Manning and Schütze (1999) propose a term-inverse document frequency measure (tf-idf) which weights each word inversely proportionally to its frequency in a document. Loughran and McDonald (2011) advocate the use of tf-idf weighting by arguing that a word's impact is likely to diminish with its frequency. Measuring the textual sentiment of annual 10-K reports with equal weights and with tf-idf weights and analyzing its impact on subsequent stock returns, they find that tf-idf weighting mitigates the impact of misclassified words in the measurement of textual sentiment. However, Henry and Leone (2016) point out that while tf-idf weighting might mitigate the impact of misclassification for frequent words, it concomitantly exacerbates the impact of misclassified words that are used only infrequently. They further argue that tf-idf weightings are sample-dependent and thus impede replication. In order to evaluate the efficacy of equal weighting versus tf-idf weighting, Henry and Leone (2016) gauge the textual sentiment in earnings announcements using both weighting schemes and analyze the subsequent capital market reaction. They find that using tf-idf weighting provides no improvement compared to equal weighting. As these issues

[15] For a comprehensive overview of studies using equal weighting, see Henry and Leone (2016).

have never been discussed for German language content analyses, we will not only measure the sentiment of CEO speeches using the context-specific BPW dictionary and compare the results to the general SENTIWS and LIWC dictionaries, but we will also evaluate the efficacy of equal weighting versus tf-idf weightings in measuring sentiment.

4. Results

4.1. Descriptive statistics

Table V-2 contains descriptive statistics for the CARs and CAVs (Panel A), for the CEO speeches' textual sentiment and other measures estimated from the CEO speeches (Panel B), as well as for the remaining control variables that we use in our multivariate regressions (Panel C).

Table V-2: Descriptive statistics

This table provides descriptive statistics for the full sample of 338 CEO speeches. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively. All variables are defined in Appendix V-1.

	Mean	Min	p25	p50	p75	Max	SD	N	T-Statistic
Panel A: CARs and CAVs									
CAR(-1,30)	-0.001	-0.277	-0.043	0.004	0.048	0.209	0.071	338	-0.202
CAR(-1,1)	0.001	-0.195	-0.017	0.000	0.017	0.095	0.029	338	0.317
CAR(2,30)	-0.001	-0.261	-0.042	-0.001	0.041	0.212	0.069	338	-0.346
CAV(-1,30)	0.999	-15.174	-4.403	-0.910	3.397	84.763	10.221	338	1.797*
CAV(-1,1)	1.502	-1.566	-0.147	0.626	1.954	19.424	3.089	338	8.942***
CAV(2,30)	-0.503	-13.948	-4.751	-1.975	1.310	83.692	8.839	338	-1.047
Panel B: CEO speeches and their sentiment									
COUNT	3,433	1,327	2,783	3,363	3,999	6,392	985	338	
IND	0.334	0.245	0.308	0.330	0.354	0.428	0.032	338	
NEG_BPW	1.154	0.235	0.759	1.057	1.508	3.237	0.549	338	
TONE_BPW	0.439	-0.207	0.268	0.459	0.621	0.894	0.237	338	
NEG_SENTIWS	1.309	0.293	0.917	1.231	1.637	2.832	0.521	338	
TONE_SENTIWS	0.740	0.420	0.670	0.754	0.824	0.947	0.105	338	
NEG_LIWC	0.359	0.000	0.233	0.337	0.460	0.962	0.182	338	
TONE_LIWC	0.717	0.213	0.649	0.741	0.815	1.000	0.140	338	
Panel C: Company-level controls variables									
EPS_SURP	0.030	-43.996	-1.567	0.374	2.055	57.060	7.933	330	
DIV_SURP_POS	0.589	0.000	0.000	1.000	1.000	1.000	0.493	338	
DIV_SURP_NEG	0.178	0.000	0.000	0.000	0.000	1.000	0.383	338	
SIZE	15,484	195	2,185	7,637	20,196	105,412	19,468	338	
M2B	2.08	0.16	1.05	1.75	2.70	10.33	1.56	338	
LV	0.07	-0.20	0.01	0.05	0.09	2.21	0.15	311	
ROA	3.68	-12.68	0.69	3.36	5.80	67.93	5.68	311	
VOLA	0.02	0.01	0.01	0.02	0.02	0.07	0.01	338	
VOLUME	28.65	0.00	2.10	6.00	31.70	406.60	53.63	337	

Panel A of Table V-2 shows that, at the mean, all CARs under investigation are economically small and not statistically different from zero. This finding indicates that, on average, we do not observe a significant market reaction around the AGM. This is in contrast to Martinez-Blasco et al. (2015), who investigate companies from the German DAX30 index and report statistically significant positive cumulative abnormal returns around the AGM. With respect to cumulative abnormal trading volumes, we find statistically significant trading volumes for CAV(-1,1), indicating that German stocks are more frequently traded around the AGM. In contrast to our finding on CARs, our results on CAVs are in line with Martinez-Blasco et al. (2015), who also report an increase in trading volume around the day of the AGM.

Panel B of Table V-2 presents summary statistics with regard to the CEO speeches and their sentiment and reveals that CEO speeches, on average, contain 1.15% negative words using the BPW dictionary and display a relative positivity, TONE_BPW, of 0.439. While the share of negative words is slightly larger using the SENTIWS dictionary (1.31%), it is much smaller employing the LIWC dictionary (0.36%). Both general dictionaries, however, also show a positive tonality. Altogether, this can be interpreted as a higher positivity than negativity of the average CEO speech. As CEOs should be expected to use public communication to present their company in a positive light, the higher positive word share does not come as a surprise. Boudt and Thewissen (2016), for instance, investigate CEO letters using the LM dictionary and find quite comparable values for negativity and relative positivity. On average, they report 1.03% of the letters' words to be negative and the relative positivity equals 0.485. Furthermore, Kim and Meschke (2014) investigate CEO interviews on CNBC using the Harvard University's General Inquirer IV-4 dictionary and find the share of negative words to be 1.38% and the relative positivity to equal 0.582. The results from the BPW word lists are hence well in line with the earlier studies.

Panel C of Table V-2 presents the control variables that we use in our multivariate regressions. Surprisingly, in only 23.3% of our observations the dividend per share is unchanged compared to the previous year, while it is decreased in 17.8% of the cases and increased in 58.9%. This is quite high compared to the results by, for example, Andres et al. (2009), who investigate German companies from 1987 to 2005 and find the dividends for German companies to be stable in 46.4%, to increase in 33.7% and to decrease in 19.9% of all cases. However, as our sample period comprises the aftermath of the financial crises, the higher fraction of dividend increases is likely to reflect stepwise re-increases of the dividend after sharp dividend cuts due to the financial crises.

Table V-3: Correlations

This table shows pairwise correlations for the full sample of 338 CEO speeches. Note that the LIWC contains word stems rather than comprehensive sets of inflections as BPW and SENTIWS. Thus, we use a stemming algorithm by Caumanns (1999) on our sample of reports before gauging the textual sentiment using the LIWC. Pearson correlations are below the diagonal, Spearman correlations are above the diagonal. P-values are in parentheses. All variables are defined in Appendix V-1.

	CAR (-1,30)	CAR (-1,1)	CAR (2,30)	CAV (-1,30)	CAV (-1,1)	CAV (2,30)	NEG_ BPW	TONE_ BPW	NEG_ SENTIWS	TONE_ SENTIWS	NEG_ LIWC	TONE_ LIWC
CAR(-1,30)		0.258	0.904	-0.001	-0.008	0.023	-0.257	0.265	-0.215	0.241	-0.207	0.200
		(0.000)	(0.000)	(0.985)	(0.879)	(0.674)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
CAR(-1,1)	0.296		-0.106	-0.090	0.031	-0.095	-0.073	0.075	-0.032	0.056	-0.038	0.059
	(0.000)		(0.051)	(0.100)	(0.573)	(0.082)	(0.179)	(0.168)	(0.561)	(0.303)	(0.489)	(0.282)
CAR(2,30)	0.912	-0.121		0.016	-0.039	0.048	-0.241	0.254	-0.211	0.230	-0.182	0.173
	(0.000)	(0.026)		(0.771)	(0.473)	(0.383)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
CAV(-1,30)	0.044	-0.077	0.078		0.713	0.949	-0.002	-0.013	0.004	-0.001	-0.058	0.005
	(0.423)	(0.159)	(0.151)		(0.000)	(0.000)	(0.975)	(0.810)	(0.946)	(0.985)	(0.291)	(0.921)
CAV(-1,1)	-0.037	0.008	-0.042	0.568		0.522	0.104	-0.129	0.091	-0.094	0.004	-0.048
	(0.504)	(0.879)	(0.447)	(0.000)		(0.000)	(0.057)	(0.018)	(0.095)	(0.084)	(0.938)	(0.383)
CAV(2,30)	0.063	-0.092	0.105	0.958	0.308		-0.044	0.038	-0.027	0.036	-0.068	0.019
	(0.246)	(0.093)	(0.054)	(0.000)	(0.000)		(0.417)	(0.484)	(0.615)	(0.513)	(0.211)	(0.724)
NEG_BPW	-0.265	-0.046	-0.255	-0.017	0.082	-0.048		-0.941	0.894	-0.880	0.692	-0.703
	(0.000)	(0.398)	(0.000)	(0.763)	(0.134)	(0.383)		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
TONE_BPW	0.274	0.060	0.259	-0.003	-0.114	0.037	-0.935		-0.857	0.904	-0.636	0.715
	(0.000)	(0.274)	(0.000)	(0.962)	(0.036)	(0.500)	(0.000)		(0.000)	(0.000)	(0.000)	(0.000)
NEG_SENTIWS	-0.221	-0.035	-0.214	-0.032	0.071	-0.062	0.901	-0.849		-0.959	0.657	-0.679
	(0.000)	(0.517)	(0.000)	(0.562)	(0.193)	(0.260)	(0.000)	(0.000)		(0.000)	(0.000)	(0.000)
TONE_SENTIWS	0.239	0.052	0.226	0.029	-0.068	0.057	-0.886	0.908	-0.950		-0.630	0.709
	(0.000)	(0.337)	(0.000)	(0.602)	(0.213)	(0.299)	(0.000)	(0.000)	(0.000)		(0.000)	(0.000)
NEG_LIWC	-0.203	-0.048	-0.190	-0.054	-0.016	-0.056	0.673	-0.619	0.662	-0.638		-0.924
	(0.000)	(0.382)	(0.000)	(0.327)	(0.777)	(0.301)	(0.000)	(0.000)	(0.000)	(0.000)		(0.000)
TONE_LIWC	0.213	0.073	0.190	0.019	-0.007	0.025	-0.675	0.691	-0.652	0.700	-0.916	
	(0.000)	(0.182)	(0.000)	(0.727)	(0.896)	(0.653)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	

Table V-3 shows the Pearson and Spearman correlations among CARs, CAVs and the measures of textual sentiment for the BPW, SENTIWS and LIWC dictionaries. For all three dictionaries, the measures of negativity appear to be negatively correlated and the measures of positivity to be positively correlated to CARs of all three time windows. However, none of the measures' correlations to CAR(-1,1) are statistically significant, while they are statistically significant at the 1%-level to CAR(2,30) and CAR(-1,30). With respect to trading volumes, the picture is less clear: The BPW and SENTIWS measures of the speeches' negativity seem to be positively correlated to CAV(-1,1) and negatively to CAV(2,30) and CAV(-1,30). The BPW and SENTIWS measures of the speeches' relative positivity seem to be negatively correlated to CAV(-1,1) and positively to CAV(2,30) and CAV(-1,30). The LIWC measures, in contrast, show no significant correlation to trading volumes.

4.2. The CEO speeches' sentiment effect on stock prices

4.2.1. The business-specific BPW dictionary

Before we proceed with the examination of the association between CEO speeches' sentiment and the stock price reaction in a multivariate analysis, we will consider the univariate dimension. In this respect, Figures V-1 and V-2 show the accumulation of abnormal returns from 5 days before to 30 days after the AGM for different levels of negativity and tonality. Figure V-1 displays the accumulated abnormal returns of high and low negativity CEO speeches, where the sample is split at the median of NEG_BPW. As can be seen from the figure, at the day of the AGM, firms with less negative CEO speeches show a 0.55% higher accumulated abnormal return than firms with more negative speeches. Over the next days after the AGM, firms with less negative CEO speeches show positive and increasing CARs. Firms with more negative CEO speeches, in contrast, display CARs that are close to zero. While the spread in CARs between the two groups increases only slowly in the first days after the AGM, it accelerates drastically from day 15 on. This may be seen as a first indication that investors indeed process qualitative information only slowly, supporting the earlier findings by Demers and Vega (2008).

Figure V-1: CARs following the AGM by high vs. low NEG_BPW

This figure shows cumulative abnormal returns (CARs) across all CEO speeches as well as segregated by a median split on NEG_BPW. The speeches' negativity and abnormal returns are estimated as described in Appendix V-1. Abnormal returns are cumulated from 5 days before the AGM until 30 days after the AGM. CARs are shown in percent.

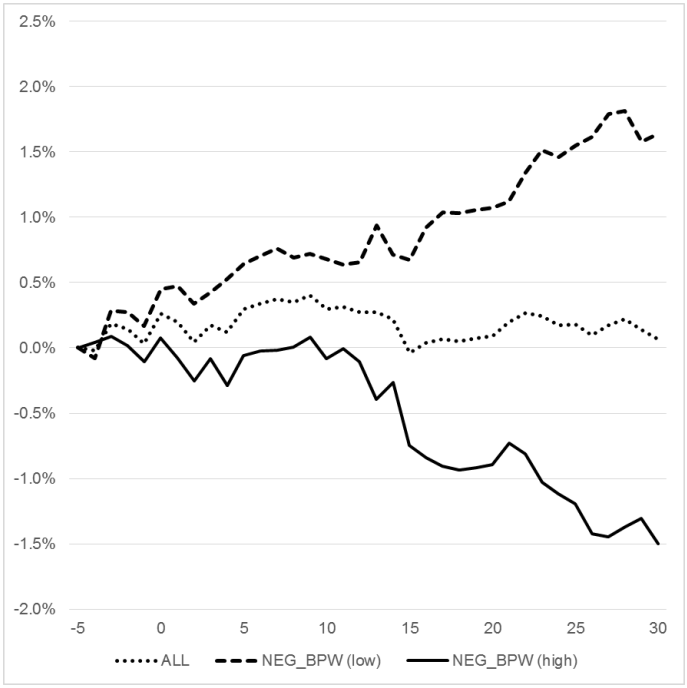


Figure V-2: CARs following the AGM by high vs. low TONE_BPW

This figure shows cumulative abnormal returns (CARs) across all CEO speeches as well as segregated by a median split on TONE_BPW. The speeches' negativity and abnormal returns are estimated as described in Appendix V-1. Abnormal returns are cumulated from 5 days before the AGM until 30 days after the AGM. CARs are shown in percent.

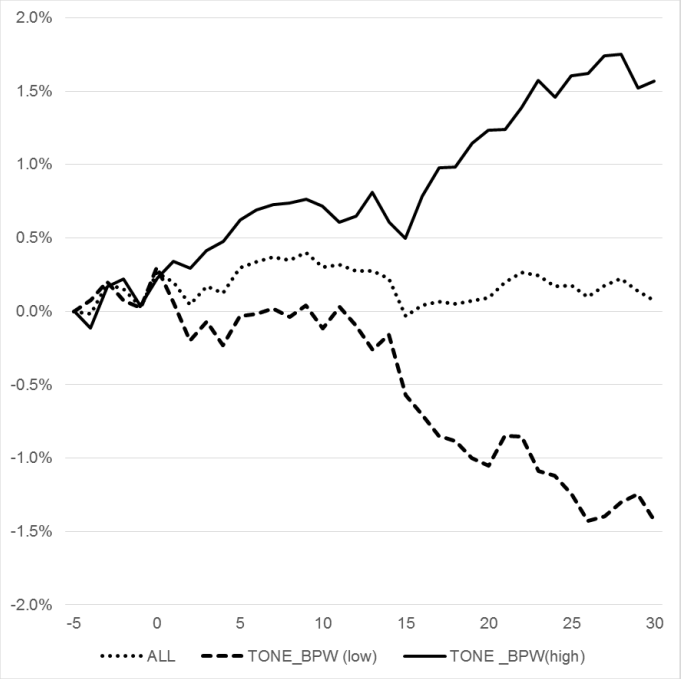


Figure V-2 depicts the development of accumulated abnormal returns, differentiating between firms with high and low tonality speeches. The sample is split along the median TONE_BPW. Similarly to the results from Figure V-1, firms with high tonality speeches display positive and increasing CARs following the AGM, while firms with low tonality speeches show CARs that are close to zero in the first days after the AGM. From day 15 on, the difference between the CARs of the two groups increases strongly as firms with low tonality speeches then show strongly negative and decreasing CARs. Again, this might be interpreted as an initial underreaction of investors to the sentiment of the CEO speeches at the AGM.

Table V-4: Test of differences of cumulative abnormal returns

This table sorts the CARs following the annual general meeting into quartiles with respect to NEG_BPW and TONE_BPW and compares the differences in mean and median CARs between the highest and lowest quartiles of textual sentiment for all time windows under investigation. Statistical significance of the differences in CARs between the highest and the lowest quartile are assessed by t and z test statistics, respectively. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively. All variables are defined in Appendix V-1.

		Q1	Q2	Q3	Q4	DIFF Q4-Q1	t-Statistic	Wilcoxon rank-sum z-Statistic
Panel A: CAR (-1,30)								
NEG_BPW	Mean	0.018	0.009	0.007	-0.038	-0.057	-5.117***	
	Median	0.019	0.008	0.008	-0.042	-0.061		-5.050***
TONE_BPW	Mean	-0.033	0.003	0.009	0.018	0.051	4.525***	
	Median	-0.034	0.006	0.008	0.017	0.052		4.441***
Panel B: CAR (-1,1)								
NEG_BPW	Mean	0.002	0.002	0.000	-0.002	-0.005	-0.980	
	Median	0.000	0.001	-0.001	-0.007	-0.008		-1.377
TONE_BPW	Mean	-0.003	0.002	0.001	0.001	0.004	0.868	
	Median	-0.005	0.000	0.000	0.000	0.005		1.077
Panel C: CAR (2,30)								
NEG_BPW	Mean	0.016	0.007	0.007	-0.036	-0.052	-4.920***	
	Median	0.016	0.012	0.009	-0.040	-0.056		-4.874***
TONE_BPW	Mean	-0.031	0.001	0.008	0.016	0.047	4.394***	
	Median	-0.036	0.007	0.010	0.016	0.052		4.451***

Table V-4 gives further information on the univariate relation between sentiment and stock market reaction. The table sorts the CEO speeches into sentiment quartiles with respect to NEG_BPW and TONE_BPW and compares mean and median CARs of the highest and lowest sentiment quartiles for all time windows. Panel A of Table V-4 presents the results for the total time period. The CAR(-1,30) differences between the highest and lowest sentiment quartile are significantly different from zero with respect to both NEG_BPW and TONE_BPW. The CAR(-1,30) mean (median) difference between the highest and lowest NEG_BPW quartiles equals -5.7 (-6.1) percentage points, and 5.1 (5.2) percentage points between the highest and lowest TONE_BPW quartiles. Panel B of Table V-4 contains the univariate results for CAR(-1,1). In this short time window around the AGM, no statistically or economically significant differences can be found between the extreme quartiles, irrespective of the sentiment measure applied. Panel C of Table V-4 presents the results for CAR(2,30). In this longer time window, we observe economically and statistically significant differences Q4-Q1 both with respect to the speeches' negativity and relative positivity.

More precisely, the CAR(2,30) mean (median) difference between the highest and lowest NEG_BPW quartiles equals -5.2 (-5.6) percentage points. With respect to TONE_BPW, the CAR(2,30) difference between the highest and lowest quartile is positive and equals 4.7 (5.2) percentage points. These first univariate results suggest that negative textual sentiment is negatively related to cumulative abnormal returns while relative positive textual sentiment shows a positive relation. Furthermore, our findings indicate that investors initially underreact to the CEO speeches' textual sentiment, as only little of the total effect is explained by an immediate reaction around the AMG.

Table V-5 finally presents multivariate regressions of CAR(-1,30), CAR(-1,1), and CAR(2,30) on NEG_BPW and TONE_BPW and a comprehensive set of control variables. Looking at the (-1,30) event window, NEG_BPW and TONE_BPW also have a strong statistically significant association with the CAR(-1,30). An increase in NEG_BPW by the interquartile change of 0.749 yields a 2.77 percentage points lower CAR(-1,30), while an increase in a CEO speech's TONE_BPW by the interquartile range of 0.353 induces a 3.14 percentage points higher CAR(-1,30). With regard to an immediate market reaction, i.e. the short-term event window (-1,1), NEG_BPW does not significantly affect the cumulative abnormal returns, thus confirming the univariate results. The relative positivity measure, TONE_BPW, in contrast, displays a statistically significant association with CAR(-1,1). However, this effect is only weakly significant and also quite small in economic terms. Nonetheless, this finding presents some first evidence that a combined positive and negative sentiment measure may capture qualitative information more effectively than a solely negative measure. In the more distant time period (2,30), NEG_BPW has a statistically significant negative effect on the cumulative abnormal returns. An increase in negativity by the interquartile change of 0.749 yields a 2.32 percentage points lower CAR(2,30). TONE_BPW also significantly affects CAR(2,30). An increase in a CEO speech's tonality by the interquartile range of 0.353 induces a 2.5 percentage points higher CAR(2,30).[16] In line with the univariate results, the mostly non-significant or only small effects of the sentiment measures in the immediate vicinity around the AGM indicate an initial investor underreaction to qualitative information as compared to the stronger reaction in the longer time period following the AGM. In this respect, our results are indeed consistent with Engelberg (2008), Demers and Vega (2008), and Price et al. (2012).

[16] Note that the CEO speeches' sentiment with respect to all our measures varies only little (Table V-2). As a consequence, interpreting the increase in terms of interquartile changes is more useful to illustrate our results.

Table V-5: Determinants of cumulative abnormal returns

This table shows regression results of CAR(-1,30), CAR(-1,1), and CAR(2,30) on our measures of textual sentiment as well as on a comprehensive set of control variables. Robust standard errors are presented in parentheses. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively. All variables are defined in Appendix V-1.

	CAR(-1,30)		CAR(-1,1)		CAR(2,30)	
	(1)	(2)	(3)	(4)	(5)	(6)
NEG_BPW	-0.037***		-0.006		-0.031***	
	(0.010)		(0.005)		(0.010)	
TONE_BPW		0.089***		0.018*		0.071***
		(0.022)		(0.010)		(0.021)
log(COUNT)	0.015	0.016	0.003	0.004	0.012	0.011
	(0.027)	(0.027)	(0.011)	(0.011)	(0.026)	(0.025)
log(IND)	0.100	0.090	0.020	0.021	0.080	0.069
	(0.084)	(0.082)	(0.034)	(0.034)	(0.084)	(0.082)
EPS_SURP	0.000	0.000	-0.000	0.000	0.000	0.000
	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
DIV_SURP_POS	-0.005	-0.005	0.001	0.001	-0.006	-0.006
	(0.011)	(0.010)	(0.004)	(0.004)	(0.011)	(0.011)
DIV_SURP_NEG	0.003	0.005	-0.004	-0.003	0.006	0.008
	(0.013)	(0.014)	(0.006)	(0.006)	(0.013)	(0.013)
log(SIZE)	0.002	0.001	0.003	0.003	-0.001	-0.002
	(0.005)	(0.005)	(0.003)	(0.003)	(0.005)	(0.005)
M2B	-0.002	-0.003	-0.004**	-0.004**	0.001	0.001
	(0.004)	(0.004)	(0.002)	(0.002)	(0.003)	(0.003)
LEVERAGE	-0.129	-0.137	-0.009	-0.009	-0.119	-0.129
	(0.088)	(0.090)	(0.050)	(0.050)	(0.092)	(0.094)
ROA	0.004	0.004	0.001	0.001	0.003	0.004
	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)
VOLATILITY	-0.565	-0.689	0.040	0.025	-0.605	-0.713
	(0.835)	(0.807)	(0.614)	(0.601)	(0.850)	(0.840)
log(VOLUME)	-0.002	-0.002	-0.002	-0.002	0.000	0.000
	(0.004)	(0.004)	(0.002)	(0.002)	(0.004)	(0.004)
Constant	0.062	-0.024	-0.004	-0.011	0.023	0.026
	(0.143)	(0.148)	(0.063)	(0.061)	(0.143)	(0.139)
Year Dummies	yes	yes	yes	yes	yes	yes
Observations	304	304	304	304	304	304
R-squared	0.140	0.142	0.051	0.056	0.133	0.131

With respect to the control variables, neither the quantity of information as measured by the speeches' length (COUNT), nor the speeches' complexity as approximated by the share of individual words (IND) are significantly associated with cumulative abnormal returns. The same is true for EPS_SURP, supporting our conjecture that any EPS surprise is likely to be already processed by financial market participants after the earlier announcement in the annual report. In contrast, a change in dividends might have an effect on the CARs, as the dividend's

payout is agreed upon at the AGM. Nevertheless, neither positive dividend surprises (DIV_SURP_POS), nor negative dividend surprises (DIV_SURP_NEG) seem to have a statistically significant effect on CARs.

To summarize, our analyses of cumulative abnormal returns highlight several interesting facts. Our measures of negative and relative positive sentiment show strong and statistically significant associations with CAR(-1,30) in univariate and multivariate analyses. When dissecting this time window into the period immediately surrounding the AGM (-1,1) and the subsequent period (2,30), we see that the market reaction occurs in a delayed fashion: Only a small fraction of the full effect is seen immediately and the larger part follows afterwards. It hence seems to be the case that the market indeed takes more time to process the qualitative information captured by the speeches' sentiment and to incorporate this in the stock price as compared to quantitative information.

Table V-6 re-estimates Table V-5, substituting NEG_BPW and TONE_BPW with the speeches' share of positive words (POS_BPW).[17] It reveals no significant relationship of POS_BPW with CAR(-1,1) or CAR(2,30). Looking at both time windows combined, we find a positive relation with CAR(-1,30) which is, however, statistically significant only at the 5% level. Compared to the speeches' negativity and relative positivity (Table V-5), the speeches' share of positive words hence seems to be less suited to capture the qualitative information of text-documents. Our results therefore support Tetlock (2007) and Loughran and McDonald (2011), who observe little incremental information using only a positive wordlist for the English language and suggest using a documents' share of negative words instead to assess its textual sentiment. We show that their observation holds for analyses on German text documents as well. Given the stronger statistical significance of the combined TONE_BPW measure, we furthermore underline the earlier suggestion that recipients tend to assess a text's positivity and negativity not in isolation but rather in relation to each other. As a consequence, tonality, i.e., relative positivity, appears to be a superior measure for capturing the qualitative information in a text or speech in the German language as well.

[17] We calculate POS_BPW analogously to NEG_BPW, where the number of negative words is replaced by the number of positive words in the respective speech.

Table V-6: Positive textual sentiment and cumulative abnormal returns

This table shows regression results of CAR(-1,30), CAR(-1,1), and CAR(2,30) on POS_BPW and on a comprehensive set of control variables. Robust standard errors are presented in parentheses. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively. All variables are defined in Appendix V-1.

	CAR(-1,30)	CAR(-1,1)	CAR(2,30)
	(1)	(2)	(3)
POS_BPW	0.014**	0.004	0.010
	(0.006)	(0.003)	(0.006)
log(COUNT)	-0.008	0.000	-0.008
	(0.025)	(0.010)	(0.024)
log(IND)	0.024	0.008	0.016
	(0.078)	(0.034)	(0.078)
EPS_SURP	0.000	-0.000	0.001
	(0.001)	(0.000)	(0.001)
DIV_SURP_POS	0.001	0.002	-0.001
	(0.011)	(0.004)	(0.011)
DIV_SURP_NEG	0.004	-0.003	0.007
	(0.014)	(0.006)	(0.014)
log(SIZE)	0.002	0.003	-0.001
	(0.005)	(0.003)	(0.006)
M2B	-0.001	-0.004**	0.002
	(0.004)	(0.002)	(0.003)
LEVERAGE	-0.182*	-0.018	-0.164*
	(0.093)	(0.050)	(0.095)
ROA	0.006*	0.001	0.005
	(0.003)	(0.002)	(0.003)
VOLATILITY	-0.778	0.007	-0.785
	(0.821)	(0.593)	(0.868)
log(VOLUME)	-0.002	-0.002	-0.000
	(0.004)	(0.002)	(0.004)
Constant	0.075	-0.006	0.081
	(0.140)	(0.060)	(0.134)
Year Dummies	yes	yes	yes
Observations	304	304	304
R-squared	0.108	0.051	0.106

4.2.2. The BPW vs. general German language dictionaries

In our analyses we so far applied the business-specific BPW dictionary. In order to evaluate its suitability for examining sentiment in business texts vis-à-vis more general word lists, we rerun our analyses using the general German language dictionaries instead. In this respect, Table V-7 re-estimates the earlier regression models using once the SENTIWS dictionary and once the LIWC dictionary to measure the sentiment of the CEO speeches. It should be noted that the following analyses employ standardized sentiment measures (with a mean of 0 and a standard deviation of 1) in order to facilitate comparisons between the

results for each dictionary. We also include the (now standardized) regression coefficients for the sentiment measured via the BPW dictionary in the first line of Table V-7.

Table V-7: Determinants of CARs: Different word lists

This table shows regression results of CAR(-1,30), CAR(-1,1), and CAR(2,30) on our measures of textual sentiment as well as on a comprehensive set of control variables. Measures of textual sentiment are standardized to have a mean of 0 and a standard deviation of 1. Robust standard errors are presented in parentheses. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively. All variables are defined in Appendix V-1.

Panel A: Regression Results for negative textual sentiment

	CAR(-1,30)			CAR(-1,1)			CAR(2,30)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
NEG_BPW	-0.020***			-0.003			-0.017***		
	(0.006)			(0.003)			(0.005)		
NEG_SENTIWS		-0.014**			-0.002			-0.011**	
		(0.006)			(0.003)			(0.005)	
NEG_LIWC			-0.01**			-0.002			-0.008*
			(0.005)			(0.002)			(0.005)
Constant	0.019	0.069	0.144	-0.004	0.003	0.016	0.023	0.066	0.128
	(0.147)	(0.146)	(0.137)	(0.063)	(0.063)	(0.058)	(0.143)	(0.143)	(0.134)
Year dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes
Controls	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	304	304	304	304	304	304	304	304	304
R-squared	0.140	0.115	0.109	0.051	0.048	0.047	0.133	0.114	0.109

Panel B: Regression results for relative positive textual sentiment

	CAR(-1,30)			CAR(-1,1)			CAR(2,30)		
	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
TONE_BPW	0.021***			0.004*			0.017***		
	(0.005)			(0.002)			(0.005)		
TONE_SENTIWS		0.015***			0.003			0.012**	
		(0.005)			(0.003)			(0.005)	
TONE_LIWC			0.010*			0.003			0.007
			(0.005)			(0.002)			(0.005)
Constant	0.015	0.08	0.128	-0.011	0.002	0.011	0.026	0.078	0.118
	(0.144)	(0.144)	(0.136)	(0.061)	(0.060)	(0.059)	(0.139)	(0.140)	(0.133)
Year dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes
Controls	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	304	304	304	304	304	304	304	304	304
R-squared	0.142	0.119	0.108	0.056	0.051	0.051	0.131	0.115	0.106

Panel A considers the negative sentiment. As can be seen, irrespective of the dictionary used, none of the measures of negative textual sentiment has a statistically significant effect on CAR(-1,1). With respect to CAR(2,30) and CAR(-1,30), in contrast, all measures show a statistically significant negative relationship. However, NEG_BPW always delivers the highest and most strongly significant coefficient. Panel B of Table V-7 refers to the tonality measure, i.e. relative positivity. In the time window (-1,30), both TONE_BPW and TONE_SENTIWS are significantly related to CARs at the 1%-level while TONE_LIWC shows a significance only at the 10%-level. Still, the effect of TONE_BPW is of higher magnitude compared to TONE_SENTIWS and TONE_LIWC. In the time window (-1,1), TONE_BPW is significantly related to CARs, while the tonality measures based on the general language dictionaries are not. In time window (2,30), only TONE_BPW and TONE_SENTIWS show a significant association with CARs, with a stronger effect again for TONE_BPW.

Table V-8 presents J-test (Davidson & MacKinnon, 1981) and Cox-Pesaran-Deaton (Pesaran & Deaton, 1978) test statistics for non-nested regressions in order to compare the in Table V-7 presented models' efficacy. The results show that none of the models using measures from the BPW dictionary can be rejected in favor of the respective models using measures from the SENTIWS or LIWC dictionaries according to both test statistics. Vice versa, the CAR(2,30) and CAR(-1,30) models including NEG_BPW (models (4) and (7)) and TONE_BPW (models (13) and (16)) are more favorable compared to the corresponding models using the general language SENTIWS and LIWC dictionaries according to both test statistics. Thus, our results indicate the superiority of context-specific dictionaries in capturing the textual sentiment of German business-related documents, underlining the earlier results from English text analyses (Price et al., 2012; Loughran & McDonald, 2011, 2015; Henry & Leone, 2016).

Table V-8: Model comparison tests

This table present J-test and Cox-Pesaran Deaton test statistics for models presented in Table V-7.

	J-test	Cox-Pesaran-Deaton test
Model (1) vs (2)	-1.08	1.00
Model (2) vs (1)	3.04***	-4.47***
Model (1) vs (3)	-0.10	0.11
Model (3) vs (1)	3.19***	-7.19***
Model (4) vs (5)	-0.24	0.23
Model (5) vs (4)	0.98	-1.36*
Model (4) vs (6)	0.07	-0.08
Model (6) vs (4)	1.10	-2.26**
Model (7) vs (8)	-1.01	0.93
Model (8) vs (7)	2.70***	-4.02***
Model (7) vs (9)	-0.14	0.14
Model (9) vs (7)	2.80***	-6.39***
Model (10) vs (11)	-0.90	0.86
Model (11) vs (10)	2.88***	-4.03***
Model (10) vs (12)	-0.16	0.16
Model (12) vs (10)	3.32***	-7.82***
Model (13) vs (14)	-0.36	0.34
Model (14) vs (13)	1.29	-1.77**
Model (13) vs (15)	0.36	-0.42
Model (15) vs (13)	1.31	-2.41***
Model (16) vs (17)	-0.77	0.73
Model (17) vs (16)	2.39**	-3.35***
Model (16) vs (18)	-0.33	0.31
Model (18) vs (16)	2.84***	-7.24***

4.2.3. Weighting schemes

The previous results have been estimated using equal weighting of words in calculating sentiment measures for the CEO speeches. In order to test whether the weighting scheme drives our results, Table V-9 re-estimates the regressions from Table V-5 using equal weighting and tf-idf weighting for calculating the sentiment measures NEG_BPW and TONE_BPW in comparison.[18]

[18] We are still employing standardized sentiment measures.

Table V-9: Determinants of CARs, by weighting schemes employed

This table shows regression results of CAR(-1,30), CAR(-1,1), and CAR(2,30) on our measures of textual sentiment as well as on a comprehensive set of control variables. Measures of textual sentiment are standardized to have a mean of 0 and a standard deviation of 1. Robust standard errors are presented in parentheses. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively. All variables are defined in Appendix V-1.

Panel A: Regression Results

	CAR(-1,30)				CAR(-1,1)				CAR(2,30)			
	equal	idf	equal	idf	equal	idf	equal	idf	equal	idf	equal	idf
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
NEG_BPW	-0.020***	-0.012**			-0.003	-0.001			-0.017***	-0.011*		
	(0.006)	(0.006)			(0.003)	(0.003)			(0.005)	(0.006)		
TONE_BPW			0.021***	0.020			0.004*	0.004			0.017***	0.016***
			(0.005)	(0.005)			(0.002)	(0.002)			(0.005)	(0.005)
Constant	0.019	-0.032	0.015	0.011	-0.004	0.003	-0.011	-0.009	0.023	-0.034	0.026	0.02
	(0.147)	(0.158)	(0.144)	(0.140)	(0.063)	(0.076)	(0.061)	(0.062)	(0.143)	(0.152)	(0.139)	(0.135)
Year dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Controls	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
N	304	304	304	304	304	304	304	304	304	304	304	304
R^2	0.140	0.110	0.142	0.140	0.051	0.046	0.056	0.054	0.133	0.112	0.131	0.131

Panel B: Model comparison tests

	J-test		Cox-Pesaran-Deaton test
Model (1) vs (2)	-0.85		0.78
Model (2) vs (1)	3.25	***	-5.85***
Model (3) vs (4)	0.69		-0.75
Model (4) vs (3)	0.91		-1.00
Model (5) vs (6)	-0.87		0.63
Model (6) vs (5)	1.59		-4.89***
Model (7) vs (8)	-0.11		0.11
Model (8) vs (7)	0.82		-0.95
Model (9) vs (10)	-0.48		0.46
Model (10) vs (9)	2.63	***	-4.45***
Model (11) vs (12)	0.77		-0.84
Model (12) vs (11)	0.57		-0.61

Panel A of Table V-9 shows that most of the coefficients estimated via tf-idf weighting are comparable in size and significance to those estimated via equal weighting. However, the coefficients of NEG_BPW in CAR(-1,30) and CAR(2,30) regressions equal -0.020 and -0.017 and are statistically significant at the 1%-level using equal weighting (models (1) and (9)), while they decrease to -0.012 and -0.011 and are only significant at the 10%-level (model (2)) and 5%-level (model (10)) with tf-idf weighting. In both cases, tf-idf weighting hence seems to unfavorably affect the results.

Panel B of Table V-9 reports the results from J-tests and Cox-Pesaran-Deaton tests. They show that none of the equally weighted NEG_BPW or TONE_BPW models can be rejected in favor of the tf-idf weighted models. Vice versa, all but two tf-idf weighted models cannot be rejected in favor of the respective equally weighted models. Only model (1) seems to be preferable compared to model (2). Model (9) seems to be preferable compared to model (10). Consequently, the results presented in Table V-9 indicate that, for our sample, tf-idf weighting seems to provide no improvement over equal weighting with respect to measures of relative positive textual sentiment. It may provide even less effective results with respect to measures of negative textual sentiment. With respect to the latter point, our results on NEG_BPW are in contrast to Loughran and McDonald (2011), who find tf-idf weighting to improve the effectiveness of their measure of negative textual sentiment. With respect to TONE_BPW, in contrast, our results are in line with Henry and Leone (2016), who find no improvement for measures of relative positivity using tf-idf weighting.

Loughran and McDonald (2011) argue that tf-idf weighting mitigates the impact of misclassified words (or noise) in the dictionaries, as words which appear more frequently are weighted less. To test this final aspect, we re-estimate Table V-7 using tf-idf weighting for all measures of textual sentiment, i.e., also those based on the SENTIWS and LIWC word lists, in Table V-10. Indeed, we find that some coefficients on general language sentiment SENTIWS and LIWC measures improve in magnitude and statistical significance. However, they still do not exceed the context-specific BPW measures. This finding is largely concordant with Henry and Leone (2016), who report that tf-idf weighting modestly increases statistical significance for general language measures of negative sentiment, but does not improve the results for measures of relative positivity.

Table V-10: Weighted CAR regressions with general language dictionaries

This table shows regression results of CAR(-1,30), CAR(-1,1), and CAR(2,30) on our measures of textual sentiment individually, as well as on a comprehensive set of control variables. Measures of textual sentiment are standardized to have a mean of 0 and a standard deviation of 1. Robust standard errors are presented in parentheses. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively. All variables are defined in Appendix V-1.

Panel A: Regression results for negative textual sentiment

	CAR(-1,30)			CAR(-1,1)			CAR(2,30)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
NEG_BPW	-0.012**			-0.001			-0.011*		
	(0.006)			(0.003)			(0.006)		
NEG_SENTIWS		-0.008			-0.001			-0.008	
		(0.006)			(0.003)			(0.006)	
NEG_LIWC			-0.008*			0.000			-0.008*
			(0.004)			(0.002)			(0.004)
Constant	-0.032	0.030	0.050	0.003	0.009	0.018	-0.034	0.021	0.031
	(0.158)	(0.159)	(0.146)	(0.076)	(0.078)	(0.067)	(0.152)	(0.157)	(0.139)
Year dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes
Controls	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	304	304	304	304	304	304	304	304	304
R-squared	0.110	0.103	0.103	0.046	0.045	0.045	0.112	0.106	0.107

Panel B: Regression results for relative positive textual sentiment

	CAR(-1,30)			CAR(-1,1)			CAR(2,30)		
	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
TONE_BPW	0.020***			0.004			0.016***		
	(0.005)			(0.002)			(0.005)		
TONE_SENTIWS		0.016***			0.002			0.014***	
		(0.005)			(0.002)			(0.005)	
TONE_LIWC			0.010**			0.001			0.009**
			(0.004)			(0.002)			(0.004)
Constant	0.011	0.053	0.082	-0.009	0.004	0.009	0.020	0.049	0.074
	(0.140)	(0.140)	(0.137)	(0.062)	(0.061)	(0.060)	(0.135)	(0.138)	(0.132)
Year dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes
Controls	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	304	304	304	304	304	304	304	304	304
R-squared	0.140	0.125	0.111	0.054	0.048	0.046	0.131	0.123	0.112

4.3. The CEO speeches' sentiment effect on trading volume

In addition to our analyses of stock prices, we also examine the relation between the CEO speeches' sentiment and the abnormal trading volume. For this examination, we employ the BPW dictionary and again start with a univariate analysis. Analogously to Table V-4 for CARs, Table V-11 shows the differences

in CAV(-1,30), CAV(-1,1), and CAV(2,30) sorted for quartiles with respect to NEG_BPW and TONE_BPW.

Table V-11: Test of differences of cumulative abnormal trading volumes

This table sorts the CAVs following the annual general meeting into quartiles with respect NEG_BPW and TONE_BPW and compares the mean and median CAV differences between the highest and lowest quartiles of textual sentiment for all time windows under investigation. Statistical significance of the CAV differences between the highest and the lowest quartile are assessed by t and z test statistics, respectively. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively. All variables are defined in Appendix V-1.

		Q1	Q2	Q3	Q4	DIFF Q4-Q1	t-Statistic	Wilcoxon rank-sum z-Statistic
Panel A: CAV (-1,30)								
NEG_BPW	Mean	0.663	1.586	1.423	0.323	-0.341	0.231	
	Median	-1.767	-0.182	-0.461	-1.274	0.493		-0.201
TONE_BPW	Mean	1.114	1.590	0.590	0.706	0.911	0.269	
	Median	-1.038	0.032	-0.999	-1.164	-0.126		-0.123
Panel B: CAV (-1,1)								
NEG_BPW	Mean	1.125	1.304	1.671	1.912	0.786	1.634	
	Median	0.377	0.556	0.853	0.735	0.358		1.481
TONE_BPW	Mean	1.853	2.250	0.877	1.033	-0.819	-1.857*	
	Median	1.076	0.924	0.320	0.327	-0.749		-1.984**
Panel C: CAV (2,30)								
NEG_BPW	Mean	-0.462	0.282	-0.248	-1.589	-1.127	-0.934	
	Median	-2.183	-1.178	-1.844	-2.982	-0.799		-0.871
TONE_BPW	Mean	-0.739	-0.660	-0.287	-0.327	0.411	0.310	
	Median	-2.362	-1.179	-2.049	-1.821	0.541		0.663

As Table V-11 shows, we find statistically significant differences in the accumulated trading volume between the fourth and first sentiment quartiles only in the short time window, CAV(-1,1), and only with respect to tonality measure. Significance is given both with parametric and non-parametric test statistics. Firms with highest tonality speeches hence show a smaller abnormal trading in the time period immediately surrounding the AGM than firms with lowest tonality speeches. This may be taken as an indication that a higher "relative negativity" seems to draw investors' attention and leads to higher abnormal trading. As we find no significant Q4-Q1 differences with respect to the longer time windows CAV(2,30) and CAV(-1,30), the observed investor attention seems to be quickly evaporating.

Table V-12: Determinants of cumulative abnormal trading volume

This table shows regression results of CAV(-1,30), CAV(-1,1), and CAV(2,30) on our measures of textual sentiment as well as on a comprehensive set of control variables. Compared to the analyses of abnormal returns, we utilize the same set of control variables for our analyses on abnormal trading volume except for log(VOLUME), which is not included in the CAV regressions. Robust standard errors are presented in parentheses. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively. All variables are defined in Appendix V-1.

	CAV(-1,30)		CAV(-1,1)		CAV(2,30)	
	(1)	(2)	(3)	(4)	(5)	(6)
NEG_BPW	-0.044		0.417		-0.461	
	(1.107)		(0.437)		(0.862)	
TONE_BPW		-2.413		-2.305**		-0.108
		(2.654)		(1.025)		(2.240)
log(COUNT)	0.917	0.056	1.520	1.063	-0.603	-1.007
	(3.033)	(3.068)	(0.942)	(0.908)	(2.573)	(2.641)
log(IND)	5.102	3.111	3.860	2.946	1.242	0.165
	(9.778)	(9.724)	(2.884)	(2.876)	(8.442)	(8.304)
EPS_SURP	0.144	0.147	0.036	0.037*	0.108	0.110
	(0.095)	(0.094)	(0.022)	(0.022)	(0.088)	(0.089)
DIV_SURP_POS	-1.429	-1.232	-0.068	0.037	-1.362	-1.269
	(1.597)	(1.573)	(0.438)	(0.432)	(1.416)	(1.386)
DIV_SURP_NEG	-0.272	-0.392	0.009	-0.077	-0.281	-0.314
	(1.674)	(1.683)	(0.647)	(0.653)	(1.280)	(1.293)
log(SIZE)	0.908***	0.944***	0.494***	0.527***	0.414	0.416
	(0.321)	(0.325)	(0.132)	(0.136)	(0.255)	(0.260)
M2B	0.273	0.375	0.250*	0.313**	0.023	0.061
	(0.402)	(0.418)	(0.138)	(0.142)	(0.330)	(0.344)
LEVERAGE	-4.190	-5.479	-2.037	-2.600	-2.153	-2.879
	(11.240)	(11.117)	(3.783)	(3.779)	(8.889)	(8.804)
ROA	-0.117	-0.075	-0.029	-0.010	-0.089	-0.065
	(0.358)	(0.351)	(0.108)	(0.108)	(0.287)	(0.281)
VOLATILITY	-240.315***	-243.972***	-99.783***	-100.202***	-140.531**	-143.771**
	(87.767)	(86.419)	(31.305)	(30.827)	(67.208)	(66.215)
Constant	-1.608	3.358	-7.746	-4.161	6.138	7.519
	(16.994)	(17.560)	(5.109)	(4.951)	(14.294)	(15.178)
Year Dummies	yes	yes	yes	yes	yes	yes
Observations	304	304	304	304	304	304
R-squared	0.168	0.170	0.255	0.267	0.104	0.103

Table V-12 is estimated analogously to Table V-5, substituting CAV for CAR. Table V-12 confirms the univariate findings from Table V-11 and shows that a higher tonality goes along with lower CAV(-1,1). Also in accordance with the univariate results, NEG_BPW does not seem to affect CAV(-1,1). With respect to the longer time horizons, we observe no statistically significant relationships between the measures of textual sentiment and CAV(-1,30) or CAV(2,30).

Table V-13 tests whether results for CAVs are influenced by the word weighting scheme applied and Table V-14 investigates the relationship among CAVs and the general language measures of textual sentiment. Similar to our

results on CARs, Table V-13 shows that tf-idf weighting does not seem to improve the results and Table V-14 reports that general language SENTIWS and LIWC measures do not possess higher explanatory power compared to the context-specific BPW measures. In particular, measuring textual sentiment via the SENTIWS or LIWC dictionaries does not yield any statistically significant relationship between textual sentiment and CAVs.

Table V-13: CAV regressions and weighting

This table shows regression results of CAV(-1,30), CAV(-1,1), and CAV(2,30) on our measures of textual sentiment individually, as well as on a comprehensive set of control variables. Measures of textual sentiment are standardized to have a mean of 0 and a standard deviation of 1. Robust standard errors are presented in parentheses. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively. All variables are defined in Appendix V-1.

Panel A: Regression results

	CAV(-1,30)				CAV(-1,1)				CAV(2,30)			
	equal	idf	equal	idf	equal	idf	equal	idf	equal	idf	equal	idf
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
NEG_BPW	-0.024	-0.346			0.229	0.106			-0.253	-0.452		
	(0.608)	(0.628)			(0.240)	(0.301)			(0.473)	(0.512)		
TONE_BPW			-0.571	-0.447			-0.546**	-0.450*			-0.025	0.003
			(0.628)	(0.607)			(0.243)	(0.242)			(0.530)	(0.510)
Constant	-1.659	-6.818	2.299	1.788	-7.265	-7.162	-5.173	-5.490	5.606	0.344	7.472	7.278
	(17.256)	(18.694)	(17.246)	(16.935)	(5.185)	(7.054)	(4.896)	(4.891)	(14.522)	(15.080)	(14.819)	(14.585)
Year dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Controls	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	304	304	304	304	304	304	304	304	304	304	304	304
R-squared	0.168	0.168	0.17	0.169	0.255	0.253	0.267	0.263	0.104	0.105	0.103	0.103

Panel B: Model comparison tests

	J-test	Cox-Pesaran-Deaton test
Model (1) vs (2)	0.65	-6.65***
Model (2) vs (1)	-0.44	0.40
Model (3) vs (4)	-0.22	0.21
Model (4) vs (3)	0.53	-0.64
Model (5) vs (6)	-0.41	0.34
Model (6) vs (5)	1.00	-2.14**
Model (7) vs (8)	-0.38	0.38
Model (8) vs (7)	1.31	-1.55*
Model (9) vs (10)	0.60	-1.05
Model (10) vs (9)	-0.14	0.13
Model (11) vs (12)	0.11	0.05
Model (12) vs (11)	0.12	-0.43

In sum, our findings on CAVs appear to some extent inverse to the results on cumulative abnormal stock returns: While the speeches' sentiment seems to be incorporated into returns rather slowly, it appears to draw investors' attention via trading volumes only during the short-term announcement period. For both returns and trading volumes, however, it is the relative positivity of the speeches

that shows the predominant effect. In the longer time periods, (-1,30) and (2,30), none of the sentiment measures displays a significant association with the CAVs. The latter finding is in contrast to Price et al. (2012), who observe for US earnings conference calls that the sentiment's effect on abnormal trading volume is statistically significant only in longer time windows. Our findings are in accordance with Martinez-Blasco et al. (2015), however, who report that the trading volume of German stocks is economically and statistically significantly increased on the day of the AGM and the two days surrounding the AGM. According to our results, this observation may at least partly be explained by the sentiment of the CEO speeches at the AGM: Speeches with particularly low relative positivity, or high "relative negativity" respectively, should draw investors' attention in the short term and go hand in hand with heightened trading volumes.

Table V-14: CAV regressions and general language dictionaries

This table shows regression results of CAV(-1,30), CAV(-1,1), and CAV(2,30) on our measures of textual sentiment individually, as well as on a comprehensive set of control variables. Measures of textual sentiment are standardized to have a mean of 0 and a standard deviation of 1. Robust standard errors are presented in parentheses. ***,**, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively. All variables are defined in Appendix V-1.

Panel A: Regression Results for negative textual sentiment

	CAV(-1,30)			CAV(-1,1)			CAV(2,30)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
NEG_BPW	-0.024			0.229			-0.253		
	(0.608)			(0.240)			(0.473)		
NEG_SENTIWS		-0.118			0.185			-0.303	
		(0.595)			(0.211)			(0.487)	
NEG_LIWC			-0.377			0.062			-0.440
			(0.485)			(0.160)			(0.401)
Constant	-1.659	-2.259	-2.077	-7.265	-7.609	-8.704*	5.606	5.350	6.627
	(17.256)	(17.157)	(16.845)	(5.185)	(5.203)	(5.169)	(14.522)	(14.442)	(14.130)
Year dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes
Controls	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	304	304	304	304	304	304	304	304	304
R-squared	0.168	0.168	0.169	0.255	0.254	0.253	0.104	0.104	0.106

Panel B: Regression Results for TONE

	CAV(-1,30)			CAV(-1,1)			CAV(2,30)		
	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
TONE_BPW	-0.571			-0.546**			-0.025		
	(0.628)			(0.243)			(0.530)		
TONE_SENTIWS		-0.329			-0.347			0.018	
		(0.597)			(0.233)			(0.474)	
TONE_LIWC			-0.097			-0.238			0.141
			(0.517)			(0.188)			(0.425)
Constant	2.299	0.168	-1.205	-5.173	-7.045	-8.083	7.472	7.213	6.878
	(17.246)	(16.948)	(16.959)	(4.896)	(5.084)	(5.146)	(14.819)	(14.385)	(14.272)
Year dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes
Controls	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	304	304	304	304	304	304	304	304	304
R-squared	0.170	0.169	0.168	0.267	0.258	0.256	0.103	0.103	0.103

5. Conclusion

CEOs' language has been repeatedly shown to exhibit information that is relevant for financial market participants, for example, in analyses on earnings conference calls (Davis et al., 2015; Doran et al., 2012; Larcker & Zakolyukina, 2012; Price et al., 2012), or CEO letters (Boudt & Thewissen, 2016). Nevertheless, CEO speeches held at companies' annual general meetings have received no attention

in studies of qualitative content analysis yet. We try to fill this gap by analyzing the investor reaction to the textual sentiment in German CEO speeches held at the companies' AGMs. We examine the speeches held by the CEOs of stock-listed German firms which regularly publish the speeches' transcripts on their internet webpages. In order to be able to analyze German texts, we utilize a novel business-specific dictionary by Bannier et al. (2017) which converts the commonly used English dictionary by Loughran and McDonald (2011) to the German language. We gather the transcripts of 338 German CEO speeches, assess the speeches' textual sentiment and measure the sentiment's effect on both stock prices and trading volumes following the AGM.

We find that the CEO speeches' textual sentiment is significantly related to abnormal stock returns and trading volume. In particular, the negativity of CEO speeches is negatively associated with abnormal returns, whereas the relative positivity of speeches is positively associated abnormal returns. With regard to the time structure of the information incorporation, we see a delayed reaction that may be interpreted as an initial underreaction to the speeches' sentiment. With respect to cumulative abnormal trading volume, in contrast, sentiment seems to have only short term effects. CEO speeches with low relative positivity are followed by increased trading volume only in the three-day window surrounding the AGM. Further, similar to content analyses on English text documents, we find that context-specific measures of textual sentiment are better suited to capture the sentiment of business-related text documents compared to general language dictionaries. Moreover, and also in accordance with literature on English content analyses (Henry & Leone, 2016), we find using combined measures of a document's positivity relative to its negativity to be advantageous compared to positive or negative measures of sentiment in isolation. Finally, our results also highlight that inverse term weighting does not yield improvements over equal weighting.

We are aware of some limitations of our analyses. First, our study is limited by the data availability of CEO speeches. As there is no compulsory register for CEO speeches, we are only able to gather CEO speeches whose transcripts are offered on the companies' homepages or sent to us on request. As most companies in our sample either offer transcripts of the speeches or do not, we can rule out the possibility that companies selectively publish only favorable speeches. However, the speeches are typically only offered a few years back, so that extending our sample poses difficulties and seems to be only possible using prospective CEO speeches. Further, the study at hand is limited to the examination of textual sentiment. Other channels of communication, for example the managers' voice, have been found to contain qualitative information as well (Hobson et al., 2012;

Mayew & Venkatachalam, 2012). Future research might extend the analysis of textual sentiment by qualitative information communicated by the managers' voice, or other channels such as, for example, gestures.

6. References

Ammann, M., & Schaub, N. (2016). Social interaction and investing: Evidence from an online social trading network. *Working Paper*.

Andres, C., Betzer, A., Goergen, M., & Renneboog, L. (2009). Dividend policy of German firms. *Journal of Empirical Finance*, 16(2), 175-187.

Arslan-Ayaydin, Ö., Boudt, K., & Thewissen, J. (2016). Managers set the tone: Equity incentives and the tone of earnings press releases. *Journal of Banking & Finance*, 72, 132-147.

Bannier, C. E., Pauls, T., & Walter, A. (2017). Content analysis of business-specific text documents: Introducing a German dictionary. *Working Paper*.

Barber, B. M., & Odean, T. (2008). All That Glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies*, 21(2), 785-818.

Blair, C., & Cole, S. R. (2002). Two-sided equivalence testing of the difference between two means. *Journal of Modern Applied Statistical Methods*, 1(1), 139-142.

Boudt, K., & Thewissen, J. (2016). Jockeying for position in CEO letters: Impression management and sentiment analytics. *Working Paper*.

Brickley, J. A. (1986). Interpreting common stock returns around proxy statement disclosures and annual shareholder meetings. *Journal of Financial & Quantitative Analysis*, 21(3), 343-349.

Caton, S., Hall, M., & Weinhardt, C. (2015). How do politicians use Facebook? An applied social observatory. *Big Data & Society*.

Caumanns, J. (1999). A fast and simple stemming algorithm for German words. Freie Universität Berlin, Fachbereich Mathematik und Informatik Ser. B, Informatik: 99-16. Berlin: Freie Univ. Fachbereich Mathematik und Informatik.

Davidson, R., & MacKinnon, J. G. (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica*, 49(3), 781-793.

Davis, A. K., Ge, W., Matsumoto, D., & Zhang, J. L. (2015). The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies*, 20(2), 639-673.

Davis, A. K., Piger, J. M., & Sedor, L. M. (2012). Beyond the numbers: Measuring the information content of earnings press release language. *Contemporary Accounting Research*, 29(3), 845-868.

Davis, A. K., & Tama-Sweet, I. (2012). Managers' use of language across alternative disclosure outlets: Earnings press releases versus MD&A. *Contemporary Accounting Research*, 29(3), 804-837.

Demers, E., & Vega, C. (2008). Soft information in earnings announcements: news or noise? *Working Paper*.

Doran, J. S., Peterson, D. R., & Price, S. M. (2012). Earnings conference call content and stock price: The case of REITs. *The Journal of Real Estate Finance and Economics*, 45(2), 402-434.

Engelberg, J. (2008). Costly Information Processing: Evidence from Earnings Announcements. *Working Paper*.

Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2008). The incremental information content of tone change in management discussion and analysis. *Working Paper*.

Ferris, S. P., Hao, Q., & Liao, M.-Y. (2013). The effect of issuer conservatism on IPO pricing and performance. *Review of Finance*, 17(3), 993-1027.

Firth, M. (1981). The relative information content of the release of financial results data by firms. *Journal of Accounting Research*, 19(2), 521-529.

García, D. (2013). Sentiment during recessions. *The Journal of Finance*, 68(3), 1267-1300.

García-Blandón, J., Martínez-Blasco, M., & González-Sabaté, L. (2012). Does the annual general meeting involve the release of relevant information in non-common law markets? Evidence from Spain. *Spanish Journal of Finance and Accounting / Revista Española de Financiación y Contabilidad*, 41(154), 209-232.

Haselmayer, M., & Jenny, M. (2016). Sentiment analysis of political communication. Combining a dictionary approach with crowdcoding. *Quality & Quantity*.

Hawkins, J. A. (2015). A comparative typology of English and German: Unifying the contrasts (1st ed., Croom Helm, London, 1986). Routledge library editions: English language: Vol. 10. London [u.a.]: Routledge.

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *Journal of Business Communication*, 45(4), 363-407.

Henry, E., & Leone, A. J. (2016). Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. *The Accounting Review*, 91(1), 153-178.

Hobson, J. L., Mayew, W. J., & Venkatachalam, M. (2012). Analyzing speech to detect financial misreporting. *Journal of Accounting Research*, 50(2), 349-392.

Huang, X., Teoh, S. H., & Zhang, Y. (2014). Tone Management. *The Accounting Review*, 89(3), 1083-1113.

Jacobi, C., Kleinen-von Königslöw, K., & Ruigrok, N. (2016). Political news in online and print newspapers. *Digital Journalism*, 4(6), 723-742.

Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3), 712-729.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33, 171-185.

Kim, Y. H., & Meschke, F. (2014). CEO Interviews on CNBC. *Working Paper*.

König, E., & Gast, V. (2012). Understanding English-German contrasts (3rd Ed.). Grundlagen der Anglistik und Amerikanistik: Vol. 29. Berlin: Schmidt.

Kothari, S. P., Li, X., & Short, J. E. (2009). The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review*, 84(5), 1639-1670.

Larcker, D. F., & Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, 50(2), 495-540.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35-65.

Loughran, T., & McDonald, B. (2015). The use of word lists in textual analysis. *Journal of Behavioral Finance*, 16(1), 1-11.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187-1230.

Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. Cambridge, Mass.: MIT Press.

Martinez-Blasco, M., Garcia-Blandon, J., & Argiles-Bosch, J. M. (2015). Does the informational role of the annual general meeting depend on a country's legal tradition? *Journal of Management & Governance*, 19(4), 849-873.

Mayew, W. J., & Venkatachalam, M. (2012). The power of voice: Managerial affective states and future firm performance. *The Journal of Finance*, 67(1), 1-44.

Mengelkamp, A., Hobert, S., & Schumann, M. (2015). Corporate credit risk analysis utilizing textual user generated content - A Twitter based feasibility study. *Working Paper*.

Olibe, K. O. (2002). The Information content of annual general meetings a price and trading volume analysis. *Journal of International Accounting, Auditing & Taxation*, 11(1), 19.

Pesaran, M. H., & Deaton, A. S. (1978). Testing non-nested nonlinear regression models. *Econometrica*, 46(3), 677-694.

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36(4), 992-1011.

Remus, R., Quasthoff, U., & Heyer, G. (2010). SentiWS - A publicly available German-language resource for sentiment analysis. LREC. 2010.

Rippington, F. A., & Taffler, R. J. (1995). The information content of firm financial disclosures. *Journal of Business Finance & Accounting*, 22(3), 345-362.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420-428.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis. MIT Press.

Tetlock, P. C. (2007). Giving content to investor sentiment. The role of media in the stock market. *The Journal of Finance*, 62(3), 1139-1168.

Tetlock, P. C., Saar-Tsechansky, M., & MacsKassy, S. (2008). More than words. Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3), 1437-1467.

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., & Kordy, H. (2008). Computergestützte quantitative Textanalyse. *Diagnostica*, 54(2), 85-98.

Zehe, A., Becker, M., Hettinger, L., Hotho, A., & Reger, I. (2016). Prediction of happy endings in German novels based on sentiment information. Proceedings of the Workshop on Interactions between *Data Mining and Natural Language Processing 2016*, 9-16.

7. Appendix

Appendix V-1: Variable descriptions

This table shows descriptions of the variables used in our analyses. COUNT, IND, and our sentiment measures are estimated directly from the CEO speeches. The data to estimate CARs, CAVs, and the remaining variables are gathered from Thompson Reuters Datastream.

Variable	Description
CAR(-1,30)	CAR(-1,30) is cumulative abnormal return from day -1 to day 30 where day 0 is the day of the AGM. Abnormal returns are estimated via a market return model as $AR_{j,t} = RI_{j,t} - RI_{CDAX,t}$ where $AR_{j,t}$ is the abnormal return for speech j at day t and $RI_{j,t}$ is the total return index for speech j at day t, which reflects the theoretical growth in value of a share over a specified period, assuming that dividends are re-invested to purchase additional units of an equity. $RI_{CDAX,t}$ is the mean total return index of the German CDAX index which 852 German stocks across the Deutsche Börse's prime and general standard.
CAR(2,30)	CAR(-1,1) is cumulative abnormal return from day -1 to day 1 where day 0 is the day of the AGM. Abnormal returns are estimated as described for CAR(-1,30).
CAR(2,30)	CAR(2,30) is cumulative abnormal return from day 2 to day 30 where day 0 is the day of the AGM. Abnormal returns are estimated as described for CAR(-1,30).
CAV(-1,30)	CAV(-1,30) is cumulative abnormal trading volume from day -1 to day 30 where day 0 is the day of the AGM. The abnormal trading volume is estimated as $AV_{j,t} = \frac{VOLUME_{j,t}}{\overline{VOLUME}_{j,t}} - 1$ where $VOLUME_{j,t}$ is the volume for company j at day t, and $\overline{VOLUME}_{j,t}$ is the mean volume for firm j from day t=-252 to t=-1.
CAV(-1,1)	CAV(-1,1) is cumulative abnormal trading volume from day -1 to day 1 where day 0 is the day of the AGM. Abnormal trading volume are estimated as described for CAV(-1,30).
CAV(2,30)	CAV(2,30) is cumulative abnormal trading volume from day 2 to day 30 where day 0 is the day of the AGM. Abnormal trading volume are estimated as described for CAV(-1,30).
COUNT	COUNT represents the CEO speeches' length in terms of the total number of words.
IND	IND is the number if individual words in a CEO speech divided by the speech's total number of words.
POS_BPW	POS_BPW represents the CEO speeche's number positive words as classified by our BPW dictionary, divided by the speech's total number of words.
NEG_BPW	NEG_BPW represents the CEO speeche's number negative words as classified by our BPW dictionary, divided by the speech's total number of words. NEG_SENTIWS and NEG_LIWC are estimated analogously using the SENTIWS and LIWC dictionary, respectively.
TONE_BPW	TONE measures a speeches positivity relative to its negativity and is calculated as $TONE_{1,j} = \frac{POSITIVE_j - NEGATIVE_j}{POSITIVE_j + NEGATIVE_j}$ where POSITIVEj is the number of positive words, NEGATIVEj the number of negative words of speech j as classified by our BPW dictionary. TONE_SENTIWS and TONE_LIWC are estimated analogously using the SENTIWS and LIWC dictionary, respectively.
EPS_SURP	EPS_SURP is the earnings surprise and is calculated as $EPS\_SURP_j = \frac{EPS_j - EPS_{j,t-1YEAR}}{STOCKPRICE_{j,t-1YEAR}} * 100$ where EPS_j is the most recent earnings per share release for the CEO's company at the time of speech j, $EPS_{j,t\text{-}1YEAR}$ is the most recent earnings per share release for the CEO's company one year before the day of speech j and $STOCKPRICE_{j,t\text{-}1YEAR}$ is the stock price of the CEO's company one year before the date of speech j.
DIV_SURP_POS	DIV_SURP_POS is a dummy variable that equals one if the dividend was increased compared to the previous year. Zero otherwise.
DIV_SURP_NEG	DIV_SURP_NEG is a dummy variable that equals one if the dividend was decreased compared to the previous year. Zero otherwise.
SIZE	SIZE measures the companies' market value at the day of the speech as the share price multiplied by the number of ordinary shares in issue. It is displayed in Euro millions.
M2B	M2B reflect the market to book ratio and is defined as the market value of the ordinary equity divided by the balance sheet value of the ordinary equity in the company.
LEVERAGE	LEVERAGE describes the total liabilities by total assets ratio.
ROA	ROA describes the companies' return on assets and is estimated as net income divided by total assets times one hundred.
VOLATILITY	VOLATILITY is estimated as the daily returns' standard deviation for the time window of minus 90 days to minus 10 days prior the AGM.
VOLUME	VOLUME describes the number of shares traded for a stock on the day of shareholder meeting and is expressed in thousands.

Affidavit

Ich erkläre hiermit, dass ich die vorgelegten und nachfolgend aufgelisteten Aufsätze selbstständig und nur mit den Hilfen angefertigt habe, die im jeweiligen Aufsatz angegeben oder zusätzlich in der nachfolgenden Liste aufgeführt sind. In der Zusammenarbeit mit den angeführten Koautoren war ich mindestens anteilig beteiligt. Bei den von mir durchgeführten und in den Aufsätzen erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis niedergelegt sind, eingehalten.

Thomas Pauls
Gießen, 29.03.2017

Submitted Papers

I. Kerl, A. G., & Pauls, T. (2014). Analyst herding and investor protection. A cross-country study. *Applied Financial Economics*, 24(8), 533-542.

II. Pauls, T., Stolper, O. A., & Walter, A. (2016). Trust and the supply side of financial advice. *Working paper.*

III. Meyll, T., Pauls, T., & Walter, A. (2017). When do households fail to repay their debt? The role of gender and financial literacy. *Working paper.*

IV. Bannier, C. E., Pauls, T., & Walter, A. (2017). Content analysis of business-specific text documents: Introducing a German dictionary. *Working paper.*

V. Bannier, C. E., Pauls, T., & Walter, A. (2017). CEO-Speeches and stock returns. *Working paper.*