The empirical relationship between forecast accuracy and recommendation profitability

Jochen Lawrenz* Klaus Schredelseker† Alex Weissensteiner‡

January 15, 2014

Abstract

This paper analyzes empirically the relation between financial analysts' forecast accuracy and their recommendation profitability and shows within a forensic approach that contrary to intuition the group of most successful recommendations is not associated with the highest accuracy on average. The finding that best performing recommendations are not the most accurate ones is even stronger under conditions of asymmetric information and weakly correlated recommendations. Our results emphasize the importance of the extent to which forecasts are correlated, and contributes to the understanding of the lack of a clear-cut relation between accuracy and recommendation profitability.

JEL classification: G10; G14; M4

Keywords: Forecast accuracy; Analysts' recommendation profitability; Information asymmetry

*Jochen Lawrenz, Department of Banking & Finance, Innsbruck University, A-6020 Innsbruck, Tel.:+43.512.507.7582, email: jochen.lawrenz@uibk.ac.at

†Klaus Schredelseker, Department of Banking & Finance, Innsbruck University, A-6020 Innsbruck, Tel.:+43.512.507.7550, email: klaus.schredelseker@uibk.ac.at

‡Alex Weissensteiner, Management Engineering, Technical University of Denmark, DK-2800 Lyngby, Tel.:+45.4525.4401, email: alwe@dtu.dk

1 Introduction

The role of financial analysts in financial markets is of immense importance and has been analyzed in numerous research papers with respect to a vast array of different aspects. We contribute to a recent strand of the literature which analyzes the association between two key characteristics of financial analysts – their forecasting abilities and their recommendation profitability. We complement existing literature by providing empirical evidence that the (contemporaneous) relation between forecast accuracy and profitability is not as straightforward as previous literature suggests. In particular, we show that on average the most profitable recommendations are *not* associated with the highest accuracy. Quite to the opposite, we document conditions where in extremis the most profitable recommendations are actually associated with the *lowest* accuracy. These findings contradict the conventional belief that more accurate forecasts lead to superior investment performance and justify further examinations of this nexus.

Our paper builds upon contributions by Loh and Mian (2006), Ertimur et al. (2007), and more recently Bonini et al. (2010), and Hall and Tacon (2010), who show empirically that on average the most accurate forecasters deliver a higher investment performance based on their recommendations than analysts with low forecasting abilities. They derive this finding basically from pairwise profitability comparisons between most and least accurate analysts (as in Loh and Mian, 2006), or by a linear (multivariate) regression approach with accuracy as an explanatory variable (as in Ertimur et al., 2007). Pairwise comparisons or linear regression by definition will not be able to detect a potential non-linear relationship between accuracy and performance. However, contributions from the information economics literature suggest that such a non-linear relationship is likely to occur in a market-based setting. Schredelseker (1984, 2001) is arguably one of the first authors who makes a strong case for a non-monotonous relationship between information accuracy and trading performance in financial markets. Building upon this reasoning, Lawrenz and Weissensteiner (2012) put forward a Bayesian learning model where financial analysts have differential forecasting abilities and show that a non-monotonous relationship between forecast accuracy and performance emerges. Crucial to their finding is the extent to which analysts' forecast errors are correlated among each other. Their results suggest that financial analysts' recommendations are likely to be more profitable if their forecasting errors are comparatively large but less correlated than by being small but highly correlated. Marinelli and Weissensteiner (2013) propose an analytical one-period framework to relate the trading profitability of each analyst to the whole information structure of the market. Compared to numerical results in previously mentioned studies, the closed-form expressions allow for a thorough comparative static analysis to investigate the role of relative information accuracy and of correlation effects. Focusing on correlated signals and/or behavior is justified and consistent with contributions which consider analysts' strategic behavior as in Fischer and Verrecchia (1998), Ottaviani and Sørensen (2006), Marinovic et al. (2013), Lamont (2002), herding behavior as in Clement and Tse (2005), or psychological traits as in Williams (2013). For either of these reasons, forecasts and forecast errors can appear correlated. The extant literature analyses correlations primarily to distinguish between the different strategic objectives of analysts,[1] and less so with respect to the impact of synchronized behavior on analysts' recommendation profitability. In our contribution, we pursue the latter implication.

Our empirical analysis contributes to the existing literature in at least four points. First, in line with the method in Loh and Mian (2006) and Hall and Tacon (2010), we form and compare portfolios (quintiles) on some sorting variable. However, we not only focus on the pairwise comparison between the highest to the lowest quintile, but we test statistically the entire structure between all quintiles in order to detect potential non-linearities. Even more importantly, we depart from Loh and Mian (2006) and Hall and Tacon (2010) by not only sorting on the accuracy measure and comparing the average performance, but by also reverting the question and sorting on the performance measure in order to compare the average accuracy. We consider such an approach as justified as it provides the answer to the natural question: What was the average forecast error of the group of most successful recommendations (most successful analysts)? Since this approach attempts to reveal the success factor of the best performing recommendations, we may label it a *forensic* approach. Interestingly, it turns out robustly that the best performing recommendations are *not* the ones associated to the

[1] See e.g. Marinovic et al. (2013), who show that reputational incentives imply more synchronized behavior in contrast to contest-like incentives.

highest forecast accuracy, and we observe a non-monotonous pattern across performance quintiles. Second, in order to elaborate on the reasons for such a pattern to occur, we double-sort on variables that can be considered as proxies for conditions which are likely candidates for explanation. These are on the one hand proxies for information asymmetry, and on the other hand proxies for the extent to which analysts' recommendations are correlated. We find that the non-monotonous pattern is even more pronounced when information asymmetries can be considered to be high, while under low information asymmetry the pattern is statistically insignificant. Likewise, the non-monotonous pattern is more pronounced when correlation is low, suggesting that low correlation compensates higher forecasting errors. In fact, we find empirically that while covariance is increasing with the forecast error, correlation is declining. This is in line with the predictions of the theoretical models of Lawrenz and Weissensteiner (2012) and Marinelli and Weissensteiner (2013), where analysts with poor forecast skills benefit from a low correlation with the other market participants. We use covariance as well as correlation as explanatory variables in the following investigations. Third, the previous findings suggest that the extent of the covariance of the analysts is detrimental to their recommendation profitability. We provide additional evidence for this interpretation by adopting the methodology from Ertimur et al. (2007) who conduct a regression approach. Our multivariate regressions yield robust negative coefficients on our proxy for covariance, confirming our hypothesis. Fourth, on the methodical side, we improve the analyses commonly used in the previous literature by providing post-hoc tests for the pairwise comparisons in the portfolio-sorting approach in order to analyze the entire structure across quintiles. For the regression approach, we address concerns with respect to heteroscedasticity and outliers by applying robust regression methods.

In general, financial analysts analyze and process valuation-relevant information. In particular they make earnings or price forecasts and issue recommendations for investment decisions. Their information processing activity is supposed to add value to their clients (see Huang et al., 2009), and more generally, to increase the information content of stock prices, see Easley et al. (1998), Barber et al. (2001), Gleason and Lee (2003) and Jegadeesh and Kim (2006). There seems to be a consensus that financial analysts have differential forecasting abilities (see e.g. Stickel, 1992; Sinha et al., 1997). However, whether higher accuracy also implies higher profitability of stock recommendations is still an open issue. On the one hand, the seminal paper by Grossman and Stiglitz (1980) implies that abnormal returns are necessary to compensate investors for their costly information-processing activities. On the other hand, in inefficient markets with information asymmetries better informed agents might exploit the ignorance of others. Therefore, conventional wisdom would suggest that the recommendations of the most accurate analyst, i.e. the one that makes the smallest forecast errors, is expected to yield the highest profitability. This wisdom is in line with empirical studies by Loh and Mian (2006) and Ertimur et al. (2007), since both papers provide empirical evidence that forecast precision and trading profitability are positively related. However, as noted by Ertimur et al. (2007), since both papers focus only on the contemporaneous relationship between accuracy and profitability, the reported abnormal excess returns among analysts cannot be considered as evidence for the existence of an implementable ex-ante trading strategy. In fact, various contributions indicate that earning abnormal trading returns based on the recommendations of financial analysts is by no means an easy task. Bradshaw (2004) shows that although earnings forecasts have the highest explanatory power for recommendations, these projections have the least association with future excess returns. Barber et al. (2001) and Mikhail et al. (2004) conclude that, after trading costs are taken into consideration, the differences in trading performance among analysts become insignificant. Mikhail et al. (1997) show that there exists a relation between analysts' experience and their forecast accuracy, but they find no corresponding relation between experience and recommendation profitability. Schredelseker (1984, 2001) provides theoretical arguments and simulation studies which suggest that the relationship between information accuracy and trading performance follows a J-shaped pattern. Building upon this reasoning, Huber (2007) and Huber et al. (2008) confirm this pattern in experimental studies. Our contribution is to the best of our knowledge the first to provide empirical evidence which adds to the understanding why clear-cut relation between accuracy and profitability is unlikely to be found in empirical data as the impact of accuracy has to be jointly analyzed with information asymmetry and correlation.

The paper is organized as follows: Section 2 introduces the research design. Section 3 shows descriptive statistics of our data. In Section 4 we present empirical results, and Section 5 concludes.

2 Research design

As our focus is on the relation between financial analysts' forecasting abilities and their associated recommendation profitability, we primarily need to operationalize these two quantities. To assess the first, i.e. the forecasting skills, the literature has used earnings per share (EPS) predictions as well as target price announcements of individual analysts. The link between both measures is, among others, discussed by Gleason et al. (2013). We follow Loh and Mian (2006) and Ertimur et al. (2007) and use EPS predictions.[2] We define the forecast error (FE) of analyst i for company j during the year y by:

$$FE_{ijy} = \frac{|Forecast_{ijy} - Actual_{jy}|}{P_{jy}}. \tag{1}$$

Analogous to Loh and Mian (2006), we consider only forecasts issued until June (cut-off month) and calculate the absolute difference between the EPS forecast and the actual EPS announced at the end of the fiscal year (see e.g. Basu and Markov, 2004, who advocate to consider the absolute instead of the quadratic forecast error). In order to allow for a comparison between the different forecasts and companies, we follow Hong and Kubik (2003) and scale the absolute forecast error in the numerator of (1) by the (unadjusted) stock price P_{jy} at the end of the corresponding year y.[3] In case analyst i issues more forecasts for company j until June of the year y, we consider the average FE_{ijy} as forecast precision for that company and year.
As the second major input, we measure the trading profitability (PFT_{ijy}) of recommendations issued by analyst i during the year y for company j. We broadly follow the approach by Barber et al. (2001) and Loh and Mian (2006), who consider average recommendations ≤ 2 as favorable, average

[2] For a discussion about the relevance of EPS forecasts in the context of trading profitability see Ertimur et al. (2007, p. 568-569).

[3] For a discussion on scaling the absolute forecast errors, see e.g. Hong and Kubik (2003), Clement (1999), Ertimur et al. (2007), Cen et al. (2013) or Williams (2013)

recommendations > 2.5 as unfavorable, and between 2 and 2.5 as neutral.[4] Loh and Mian (2006) form corresponding long and short portfolios for favorable and unfavorable recommendations, respectively, for each quintile. Thus, for strong buy and buy recommendations we take a long position, while for underperformance and sell recommendations we take a short position in the stock.[5] Finally, PFT_{ijy} is then defined as the average excess return of this trading strategy (i.e. tradings based on the single recommendations) over the US-Libor rate for a holding period of 6 months (as in e.g. Martinez, 2011). In order to account for dividend payments and stock splits in the calculation of PFT_{ijy} we use adjusted stock prices.

As discussed in the introductory section, we intend to augment the analysis of the accuracy-profitability relation by the extent to which the forecasts are synchronized, i.e. correlated among analysts. Since we consider individual forecast errors and recommendation trading returns, a covariance or correlation obviously cannot be determined for single (non-overlapping) observations. Thus, we need a proxy for this quantity, which we define as the average covariance COV_{ijy} and the average correlation COR_{ijy} between the trading strategy and the overall market by using daily log-returns of the recommended stock. Note that the covariance between returns of one stock r_j and the overall index r_{IDX} can be written as sum of the covariances between j and all components d of the index:

$$\operatorname{cov}(r_j, r_{IDX}) = \operatorname{cov}\left(r_j, \sum_d w_d r_d\right) = \sum_d w_d \operatorname{cov}(r_j, r_d), \qquad (2)$$

with w_d the corresponding weights. Therefore, COV_{ijy} and COR_{ijy} can be interpreted as proxies for the aggregated average return co-movement between trading recommendations of analyst i (following company j) with the entirety of other analysts during the year y. If recommendations have a potential price impact as documented in Stickel (1995) and Brav and Lehavy (2003), then an increasing covariance between the analysts is supposed to

[4] Hall and Tacon (2010) consider ratings ≤ 1.75 as favorable and those ≥ 2.75 as unfavorable.

[5] Since we consider individual recommendations in our study which belong to one of the five I/B/E/S categories, we depart slightly from Loh and Mian (2006) in that we cannot split the "hold" category.

reduce the profitability. This effect was found in a theoretical model by Lawrenz and Weissensteiner (2012). For the possible impact of the correlation on the expected trading payoff of financial analysts and the role of relative accuracy see Marinelli and Weissensteiner (2013).
Note that while the approach in Loh and Mian (2006) and Hall and Tacon (2010) is to first cluster analysts according to their EPS forecast precision into 5 and 3 categories respectively, then calculate the average recommendation of each category and compare the corresponding trading performance of these different categories, we follow Ertimur et al. (2007) and base our analysis on single FE_{ijy}, PFT_{ijy} combinations in order to avoid that due to the clustering relevant information is lost or smoothed out.

To analyze the relation between accuracy and profitability, we use different techniques. First, in line with e.g. Loh and Mian (2006) or Hall and Tacon (2010) we form portfolios of observations by some sorting variable. The portfolio formation needs some deliberate choices with respect to the number of categories and how to determine cut-off points. Following the bulk of the literature, we report results for five categories, which turns out to strike a reasonable balance between sufficient structure and statistical significance.[6] To determine the cut-off points for categories, two approaches are conceivable. Either we determine categories with an equal number of observations, i.e. quintiles, or we determine categories by forming equal intervals in the range of the sorting variable.[7] For most analyses, we report results for both approaches.
To assess the statistical difference between categories, we not only consider the difference between the lowest and highest quintile and test for a significant deviation from zero as in the previous literature. Since we focus on the entire pattern across (multiple) categories, we first perform an ANOVA test to determine if categories differ significantly. Since diagnostics indicate that assumptions for a parametric ANOVA are likely to be violated, we use the non-parametric Kruskal-Wallis test (i.e. a generalization of the Mann–Whitney U test for multiple categories). Rejecting the null of equal

[6] In unreported results, we find similar relations for 3, 7, and 10 categories.
[7] Given that the sorting variables are usually far from being uniformly distributed, forming quintiles has the drawback that category intervals in the sorting variable are highly unequal, while forming equally spaced intervals yields categories with unequal number of observations.

distributions across categories does not yet enable a statistical assessment of the pattern across categories. Thus, we perform post-hoc tests to determine which categories differ from each other significantly. A variety of post-hoc tests exist which basically propose different procedures for how to adjust for the increased chances to commit type I-errors in multiple pairwise comparisons, from which we use the procedure suggested by Dunn (1964).

Second, besides the portfolio-sorting approach, we further analyze our hypotheses by using a (multivariate) regression approach, as e.g. in Ertimur et al. (2007). Using a second method is a way to verify whether our results are robust to different statistical techniques. We apply standard OLS methods (with heteroscedasticity-adjusted standard errors). Furthermore, as diagnostics suggest that OLS assumptions are likely to be violated, and in an attempt to cope with the impact of extreme values in the data (i.e. outliers and leverage points which might distort the parameter estimation), we use the robust regression method to estimate the relation between accuracy and profitability. We choose the so-called MM-estimation which addresses the impact of outliers in both the independent and the dependent variable.[8] For most analyses, we report results from both regression methods.

3 Data

Our analysis is based on the constituents of the S&P500 index. We collect data on financial analysts' earnings forecasts and the actual earnings over time through I/B/E/S Detail File. Analysts' stock recommendations are extracted from the I/B/E/S Detail Recommendations File. The recommendations in the I/B/E/S database are classified according to the usual 5 point scale from 1 (strong buy) to 5 (sell). If analysts provide recommendations on a different scale, these are mapped by I/B/E/S onto the 5 point scale. Furthermore, we use unadjusted and adjusted (for stock splits and dividends) daily stock prices, which we obtain from Thomson Reuters Datastream.

The I/B/E/S datafile records earnings information from 1976 onwards, however recommendation information is only available beginning with late 1993. Thus, our analysis starts with the year 1994, for which joint earnings and

[8] Robust regression is performed by using Eviews which uses the bisquare objective function and the Huber Type I for computing the coefficient covariance matrix by default.

recommendation information is available for the first year. The end of our observation sample is 2011, thus our sample covers a rather extensive 18-year period from 1994 to 2011 including stock markets up- as well as downswings. This sample includes 25,316 observations. We apply a couple of filter rules to the initial sample. First, in order to avoid the impact of extreme outlier observations, we drop observations below (above) the 1%- (99%-)quantile of our performance measure PFT_{ijy}.[9] Furthermore, to avoid outliers in the forecast accuracy variable FE_{ijy}, we remove observations above the 99%-quantile of FE_{ijy}.[10] Together the removal of outliers reduces the sample size to 24,561. Second, following e.g. Ertimur et al. (2007) and Loh and Mian (2006), we require a minimum number of observations for each analyst, which we choose to be 5. This filter is intended to ensure that only observations for sufficiently active analysts are considered. The removal of inactive analysts' observations reduces the sample to 18,637. We take this sample to be our base sample. Descriptive statistics for the base sample are collected in Table 1 on a yearly basis. The second column reports the number of (accuracy-recommendation profitability) observations N, for which an increasing trend up to 2007 can be observed with a stagnation thereafter. The third column reports the mean of our measure of forecast accuracy ($\overline{FE}$) which assumes a value of 0.0087 over the whole sample. Abnormally high values of $\overline{FE}$ occur in the years 2001 and 2008-10. The next (fourth) column (SFE) shows the standard deviation of FE in order to indicate the dispersion of forecast errors. Again, high values occur in 2001/02 as well as in 2007-09. Together, this hints towards increased forecast uncertainty among financial analysts around the time of market turmoils. Our results for the mean and standard deviation of FE are consistent with previous literature, as e.g. in Loh and Mian (2006) who find a mean of 0.0126, a standard deviation of 0.0203, and a maximum of 0.246.[11] The fifth and sixth column report the mean ($\overline{PFT}$) and the standard deviation ($SPFT$) of our recommendation performance measure. Again, we see particularly negative values in times of market turmoil around 2000/01 and 2007/08. In particular, the highest volatility is observed in 2008. Interestingly, the average performance

[9] In the initial sample, the min and max of PFT_{ijy} were -3.55, and 2.30 respectively. After removing the bottom and top 1%, the min and max are -0.95 and 0.69 respectively.

[10] In the initial sample, the max of FE_{ijy} was 27.7. After removing the top 1%, the max is 0.198.

[11] See Table 4 in (Loh and Mian, 2006, p. 469).

Table 1: Descriptive statistics

This table reports descriptive statistics for the full sample. N is the number of observations, $\overline{FE}$ is the mean absolute forecast error, SFE the standard deviation of the forecast error. $\overline{PFT}$ reports the mean performance based on the recommendations, $SPFT$ is the standard deviation of the performance. I and J report the number of analysts and companies. The tenth and eleventh column show the mean and the maximum number of analysts per company ($\bar{I}_j$, $\max I_j$), while the last two columns report the mean and maximum number of observations per analyst ($\bar{N}_i$, $\max N_i$). The last row gives descriptive statistics for the full sample.

	N	$\overline{FE}$	SFE	$\overline{PFT}$	$SPFT$	I	J	$\bar{I}_j$	$\max I_j$	$\bar{N}_i$	$\max N_i$
1994	441	.0045	.0089	.0082	.180	243	198	2.22	8	1.81	10
1995	386	.0052	.0112	.0579	.204	240	184	2.09	9	1.60	7
1996	694	.0056	.0115	.0424	.230	356	225	3.08	16	1.94	9
1997	679	.0059	.0139	.0478	.245	378	236	2.87	15	1.79	12
1998	714	.0081	.0183	.0158	.284	405	242	2.95	12	1.76	18
1999	751	.0068	.0150	.0067	.296	421	253	2.96	10	1.78	16
2000	1118	.0070	.0145	-.0391	.299	556	277	4.03	19	2.01	12
2001	1145	.0126	.0226	-.0523	.258	592	294	3.89	16	1.93	9
2002	1655	.0090	.0217	-.0688	.230	731	311	5.32	20	2.26	11
2003	1286	.0061	.0123	.0758	.213	650	311	4.13	15	1.97	8
2004	1314	.0068	.0128	.0081	.185	666	313	4.19	21	1.97	12
2005	1196	.0071	.0107	.0274	.184	623	322	3.71	15	1.91	11
2006	1303	.0076	.0133	.0035	.163	662	322	4.04	19	1.96	13
2007	1425	.0086	.0208	-.0254	.217	658	318	4.48	19	2.16	13
2008	1412	.0161	.0273	-.1324	.316	634	313	4.51	15	2.22	12
2009	1280	.0130	.0235	.0685	.265	540	324	3.95	14	2.37	14
2010	1295	.0104	.0170	.0602	.190	538	332	3.90	17	2.41	19
2011	543	.0029	.0034	.0240	.161	300	64	8.48	27	1.81	7
Total	18637	.0087	.0180	-.0006	.242	1679	394	47.3	209	11.1	76

over the full sample is effectively zero.[12] Taken at face value, this means that the average investment performance on the basis of financial analysts' recommendations as entire profession does not generate any excess profitability over the observed period.

Column seven and eight report the number of analysts (I) and the number of companies (J). The full sample includes a total of 1679 analysts, as well as 394 companies from the S&P500 for which our filter rules yield valid observations. The next two columns report the mean ($\bar{I}_j$) and the maximum ($\max I_j$) of the number of analysts per company. In the full sample, we find an average of 47 analysts following a given company, with a maximum of

[12] The point estimate is even slightly negative at -.0006 as indicated in the last row, but given the dispersion certainly not statistically different from zero.

209 analysts for the most-covered company. The last two columns report the mean ($\bar{N}_i$) and maximum number of observations ($\max N_i$) per analyst. On average, our sample includes 11 observation for each analyst, with a maximum number of 76 observations for the most-active analyst.
In comparison to previous studies, our base sample is smaller although we cover a longer time period. Loh and Mian (2006) and Ertimur et al. (2007) use a sample size of 32,147 and 64,206 respectively. The reason is that we restrict ourselves to S&P500 companies which we consider to be more liquid, and that we require both an earnings estimate as well as a recommendation for any given company in any given year.

4 Empirical results

4.1 Evidence from previous literature

In this section, we first investigate if we are able to replicate with our method and sample similar results as in the previous literature. As discussed in the introductory section, the consensus so far seems to be that more accurate analysts are able to issue recommendations which yield superior investment results than recommendations from less accurate analysts (see e.g. Ertimur et al., 2007; Loh and Mian, 2006). To make our results more comparable to those of Loh and Mian (2006), we also average the accuracy measure FE_{ijy} over the different individual analysts i to get FE_i. For FE_i we form quintiles to obtain five categories with category $FE1$ ($FE5$) representing analysts with the smallest (highest) average forecast error. For the observations in these quintiles, we adopt two approaches to provide robust results. First, we use all observations on the disaggregated level (FE). Second, we average the observations of the individual analysts' (FE_i). In the second case, the mean performance is calculated on the basis of average individual analysts performance. Obviously, this reduces the number of observations to the number of analysts, and yields different results if the number of observations per analyst is not equal. Results for the disaggregated and aggregated approach are reported in Table 2 on the left and right side respectively.

Panel A in Table 2 reports the mean and the standard deviation for the performance measure ($\overline{PFT}$ and $SPFT$) as well as for the forecast error ($\overline{FE}$ and FE). By construction, the mean forecast error increases over categories.

Table 2: Previous literature

This table reports results for the full sample $N = 18637$. In a first step we form quintiles based on the forecast precision of the single analyst i. The left side shows the results on a disaggregated level. The results on the right side is based on the average observation of each analyst. Panel A reports average profitability ($\overline{PFT}$) with standard deviation ($SPFT$), average forecast error ($\overline{FE}$) with standard deviation SFE, as well as average covariance ($\overline{COV}$) and average correlation ($\overline{COR}$). Panel B compares $\overline{PFT}$ of category 1–4 against that of category 5. Panel C performs the Kruskal-Wallis and Panel D gives the corresponding post-hoc tests. Panel E reports results for an univariate regression analysis, where trading profitability is explained by forecast precision.

	FE					FE_i				
$FE-Cat$	1 (low)	2	3	4	5 (high)	1 (low)	2	3	4	5 (high)
A. Estimates										
$\overline{PFT}$	.0164	.0012	-.0033	-.0023	-.0124	.0221	.0013	-.0012	-.0056	-.0206
$SPFT$	.224	.215	.253	.249	.261	.094	.091	.098	.085	.112
$\overline{FE}$	.0015	.0032	.0058	.0098	.0214	.0014	.0032	.0058	.0098	.0233
SFE	.0018	.0042	.0076	.0148	.0319	.0005	.0005	.0009	.0015	.0115
$\overline{COV}$	.000129	.000115	.000141	.000136	.000144	.000130	.000115	.000138	.000137	.000145
$\overline{COR}$	.393	.370	.365	.352	.329	.390	.377	.361	.351	.330
N	3305	3427	3740	4350	3815	336	336	336	336	335
B. t-Test										
$FE5$	5.01**	2.43*	1.53	1.77	–	5.33**	2.78**	2.38*	1.95	–
	(.000)	(.015)	(.125)	(.076)		(.000)	(.005)	(.017)	(.051)	
C. Kruskal-Wallis										
	N	df	F	p		N	df	F	p	
$\overline{PFT}$	18637	4	19.4**	0.001		1679	4	22.8**	0.000	
D. Post-hoc tests										
1		3.26*	2.97*	2.92*	4.19**		2.48	2.65	3.59**	4.52**
		(.011)	(.030)	(.034)	(.000)		(.131)	(.079)	(.003)	(.000)
2			0.36	0.52	0.85			0.17	1.10	2.04
			(1.0)	(1.0)	(1.0)			(1.0)	(1.0)	(.411)
3				0.15	1.24				0.93	1.86
				(1.0)	(1.0)				(1.0)	(.621)
4					1.44					0.93
					(1.0)					(1.0)
E. Regression results										
	$PFT = c + b\,FE + \epsilon$					$PFT_i = c + b\,FE_i + \epsilon$				
c	.033**	.015**	.024**	.022**	.012*	.066**	-.018	.046	.021	.014
	(.000)	(.001)	(.000)	(.000)	(.020)	(.000)	(.479)	(.181)	(.506)	(.220)
b	-11.2**	-4.48**	-4.80**	-2.55**	-1.16**	-30.2**	5.93	-8.24	-2.75	-1.49**
	(.000)	(.000)	(.000)	(.000)	(.000)	(.003)	(.432)	(.168)	(.404)	(.005)
R^2	.008	.007	.020	.022	.019	.025	.001	.003	.002	.020
	Total sample c: .012**(.000), b: -1.55**(.000), R^2: 0.013					Total sample c: .012**(.000), b: -1.56**(.000), R^2: 0.022				

The same holds also for the corresponding volatility. More interestingly, the numerical values for the performance measure decline from categories one to

five, suggesting that the average performance of more accurate analysts is substantially higher than less accurate analyst, the difference amounting to 0.0288 for the disaggregated data, and 0.0427 on the basis of analysts' mean performance.[13] Futhermore, we observe that, in line with predictions of the models by Lawrenz and Weissensteiner (2012) and Marinelli and Weissensteiner (2013), the average correlation is lower for analysts with a low forecast precision, while covariance tends to be high.
The previous literature basically compares the performance of the most accurate quintile against the least accurate quintile and tests if means are significantly different from each other. We provide results of the corresponding t-tests in panel B, where we compare the mean performance of category five ($\overline{PFT}5$ least accurate analysts) against that of categories one to four. The t-statistic for comparing $\overline{PFT}5$ to $\overline{PFT}1$ (most accurate quintile) is 5.01 and therefore strongly significant, confirming the results as in e.g. Loh and Mian (2006) that more accurate analysts seem to provide more profitable recommendations. However, strictly speaking this result does not yet establish the more general claim that the relationship between accuracy and profitability is monotonically increasing as this is suggested by the numerical values of the mean performance across accuracy quintiles. To corroborate this claim, one might use the additional t-test between $\overline{FE}5$ and categories two to four. Results in panel B seem to confirm the claim, as nearly all estimates are significant at least at the 10% level. As discussed in Section 2, we consider Kruskal-Wallis and associated post-hoc tests as appropriate statistical method, for which we report results in panel C and D. The first row shows strongly significant F-values, allowing us to reject the hypothesis that all categories are drawn from the same distribution. The corresponding post-hoc tests show which of the categories differ from each other significantly in pair-wise comparisons. In line with the simple t-test, we still have a strongly significant difference between $\overline{FE}1$ and $\overline{FE}5$. However, none of the other categories is even close to being statistically different from category $\overline{FE}5$ in contrast to the results from the t-test. Actually, the results suggest that only $\overline{FE}1$ is different from the other categories, and none of

[13] As we consider a holding period of 6 months, we find monthly values of 0.48% and 0.71% respectively, which is of a similar magnitude as in Loh and Mian (2006, p. 474)

the other categories is different from each other.[14] From this, we are forced to conclude that we cannot establish the claim that performance generally increases with accuracy. Although the most accurate forecasts are found to yield more profitable recommendations than less accurate analysts, we do not see any statistical differences among the remaining four accuracy quintiles.

To further elaborate on this finding, we follow a second, methodologically different approach, where we conduct a univariate regression analysis with performance as response and forecast error as predictor variable, i.e. we estimate the model $PFT = c + b\,FE + \epsilon$ and $PFT_i = c + b\,FE_i + \epsilon$. We perform the regression for the total sample as well as for each accuracy quintile separately and report results in panel E of Table 2. For the total sample, we find a strongly significant negative coefficient of -1.55 (-1.56), suggesting a negative relationship. However, R^2 are modest at 0.013 (0.022). At the level of accuracy quintiles, we find a more differentiated result, in the sense that in the most accurate quintile coefficients are -11.2 and -30.2 respectively, while being much lower in absolute terms in all other quintiles. Even more interesting, for the mean performance sample, coefficients are insignificant for the middle categories $\overline{FE}2$ to $\overline{FE}4$. Taken together, regression results further confirm the finding that a clear negative relationship between forecast error and performance can only be established for the quintile of the most accurate analysts, while we only find weak or no relationship in the remaining four accuracy quintiles.
Thus, the evidence from replicating results from the previous literature suggests that the average recommendation performance of the most accurate quintile is significantly different (and higher) from all other quintiles, but that we cannot confirm the more general claim that there exists a general negative relationship between forecast errors and profitability as we do not find any significant differences among accuracy quintiles apart from the most accurate one. This finding leads us to investigate more closely the relationship between accuracy and profitability in the next section.

[14] For the average observations per analyst (right side), only two differences are significant, namely between category 1 and 4, and between 1 and 5.

4.2 Relationship between accuracy and profitability

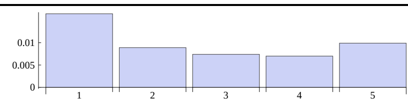
Here we adopt a different approach as in the previous section, where we form quintiles on the basis of the accuracy measure and compare mean profitability across accuracy quintiles. This sorting approach basically answers the question: Which average performance can be expected from analysts classified in a given accuracy quintile? Or in other terms: What kind of performance is associated to a given accuracy? The approach taken in this section reverts the question, and asks: What kind of accuracy is associated to a given performance? Thus, we sort our data on the basis of performance (rather than accuracy) and investigate which characteristics explain the given performance. Such an approach may be considered as a forensic analysis, since we are interested in the characteristics – in particular in the accuracy – of the (ex post) successful analysts (or recommendations).

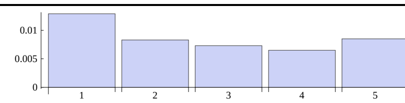
To construct performance categories, we sort our profitability measure PFT in ascending order and form five categories. Two ways how to define categories are conceivable: Either in terms of intervals with the same distance (D) in the performance measure, or in terms of an equal number of observations (N). We provide results for both approaches, where the data in Table 3 and subsequent tables are arranged such that results on the basis of equal intervals are on the left side and results on the basis of the same number of observations are on the right side of the table. In panel A of Table 3, we report descriptive statistics for the forecast error across performance categories 1 (lowest) to 5 (highest profitability). The first row ($\overline{FE}$) contains the average forecast error in each category, whose numerical values are shown graphically in the barcharts above. Contrary to intuition, which would suggest from results of the previous section that most successful analysts might display the highest accuracy, we find a different pattern. While being in line with intuition, the least successful recommendations are associated with the highest forecast error. However, the second largest forecast errors occur in the category 5 of the most profitable recommendations. Thus, we observe a non-monotonous relationship between accuracy and profitability. The finding of this non-monotonicity is, e.g., in line with the theoretical predictions of Schredelseker (1984, 2001) as well as with the model put forward in Lawrenz and Weissensteiner (2012). To corroborate the results statistically, we perform the Kruskal-Wallis test for which results are reported

Table 3: Forensic analysis

This table reports results for the full sample $N = 18637$. The left side shows results for 5 performance categories classified according to equal intervals ($PFT - D$), while the right side shows results for equal quantiles ($PFT - N$). Panel A reports estimates for the mean forecast error ($\overline{FE}$) with the standard deviation (SFE), mean profitability ($\overline{PFT}$) with standard deviation ($SPFT$), together with number of observations (N). Panel B reports results of the Kruskal-Wallis one-way ANOVA to test for equality of distribution. To further analyze between which categories differences are significant, Panel C reports Post-hoc tests, where pairwise multiple comparisons are done on the basis of the procedure of Dunn (1964) to adjust for familywise type I error. p-values are below statistics in parantheses. Finally, panel D reports standard t-test results for a comparison of category 5 against category 2, 3, and 4. F reports the Levene-test-statistic for equal variance, while T is the corresponding test-statistic (df reports degree of freedoms).

A. Descriptive statistics





	$PFT-D$					$PFT-N$				
	1 (low)	2	3	4	5 (high)	1 (low)	2	3	4	5 (high)
$\overline{FE}$	0.0165	0.0089	0.0074	0.0070	0.0099	0.0129	0.0083	0.0073	0.0065	0.0085
SFE	0.0285	0.0173	0.0160	0.0138	0.0203	0.0238	0.0169	0.0170	0.0125	0.0171
$\overline{PFT}$	-0.486	-0.164	0.003	0.181	0.418	-0.354	-0.095	0.015	0.119	0.312
$SPFT$	0.164	0.055	0.056	0.054	0.096	0.175	0.038	0.029	0.032	0.115
N	1837	3501	7110	4581	1608	3728	3727	3728	3727	3727

B. Kruskal-Wallis

	N	df	F	p	N	df	F	p
$\overline{FE}$	18637	4	383.3^{**}	0.000	18637	4	353.4^{**}	0.000

C. Post-hoc tests

	1	2	3	4	5	1	2	3	4	5
1		10.19^{**}	18.25^{**}	16.87^{**}	10.71^{**}		10.89^{**}	17.04^{**}	15.34^{**}	11.55^{**}
		(.000)	(.000)	(.000)	(.000)		(.000)	(.000)	(.000)	(.000)
2			8.08^{**}	6.91^{**}	-1.82			6.14^{**}	4.44^{**}	-.51
			(.000)	(.000)	(.677)			(.000)	(.000)	(1.0)
3				-.62	-4.05^{**}				0.08	-5.48^{**}
				(1.0)	(.001)				(.895)	(.000)
4					-3.45^{**}					-3.78^{**}
					(.005)					(.002)

D. t-test against $\overline{FE}5$

	F	p	df	T	p	F	p	df	T	p
2	15.5	.000	2718	-1.78	.076	1.17	.279	7452	-.68	.494
3	60.7	.000	2082	-4.75^{**}	.000	9.59	.002	7452	-3.08^{**}	.002
4	100.1	.000	2156	-5.45^{**}	.000	68.3	.000	6838	-5.65^{**}	.000

in panel B, and which clearly rejects the hypothesis that all categories are drawn from an equal distribution. However, only post-hoc tests are able to confirm the interpretation of the visual inspection, and we report results in panel C. Results show that categories are significantly different from each other with two notable exceptions: Categories 3 and 4, and categories 2 and 5 do not differ significantly. The most interesting results are observed for the pair-wise comparisons of category 5 against the other categories. Except for category 2, differences are significant, which confirms the interpretation that most successful recommendations are associated with a statistically higher forecast error than in category 3 and 4, and which is indistinguishable from the forecast error of category 2. Results are qualitatively the same for both approaches to form categories. Although we consider Kruskal-Wallis post-hoc test to be the superior methodological approach, panel D reports for robustness reasons simple t-tests of category 5 against 2 to 4, which basically confirm the same interpretation. Thus, the statistical evidence confirms that the non-monotonous pattern is significant.

4.3 Information asymmetry

4.3.1 Industry sorting

In order to better understand the observed non-monotonous pattern, we provide further analyses under which conditions this pattern emerges or is particularly pronounced. A first natural choice is to sort our sample according to industries. Different previous papers document that there are significant differences across industries in terms of analyst following, see, e.g., Bhushan (1989). Thus, we use the 5-industry classification of K. French[15] to split our sample. Our focus is on the detection of the non-monotonous pattern, therefore in the interest of available space, we only report the Kruskal-Wallis and post-hoc tests for the pair-wise comparison between category 5 (highest PFT) and 4, and between 5 and 3. Results are reported in Table 4. Columns 2 to 5 give descriptive statistics for industries, while columns 6 & 7 (8 & 9) report the differences in FE between category 5 and 4 (5vs4) and 5 and 3 (5vs3) for the equal interval (quintiles) categories. For each industry, we report the difference, the standardized test statistic, and the p-value. We ob-

[15] See `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

Table 4: Financial industries

This table shows results for different sectors of the full sample $N = 18637$ using the 5-industry classification of K. French. In addition to indicate the total number of observations N, the average forecast precision $\overline{FE}$, the corresponding standard deviation SFE and the average correlation $\overline{COR}$ with the market, we classify the observations in each industry according to equal performance intervals ($PFT - D$) and equal performance quantiles ($PFT - N$). For each type of categorization, we report the difference in FE, the standardized test statistic and the p-value between the category 5 (highest profitability) and group 4 and group 3 (5vs4 and 5vs3).

					$PFT-D$		$PFT-N$	
	N	$\overline{FE}$	SFE	$\overline{COR}$	5vs4	5vs3	5vs4	5vs3
Ind1	3896	.0041	.0121	.3526	.00039	.00012	.00043	.00012
Cnsmr					.630	.167	.189	.991
					(1.0)	(1.0)	(1.0)	(1.0)
Ind2	5367	.0115	.0180	.3403	.00475	.00453	.00308	.00282
Manuf					6.06^{**}	7.05^{**}	6.06^{**}	7.12^{**}
					(.000)	(.000)	(.000)	(.000)
Ind3	4787	.0083	.0179	.4060	.00242	.00073	.00099	-.00016
Hitec					1.47	3.45^{**}	2.42	2.56
					(1.0)	(.005)	(.152)	(.105)
Ind4	1419	.0031	.0069	.3302	.00107	.00056	.00029	.00031
Health					.225	.684	.64	.03
					(1.0)	(1.0)	(1.0)	(1.0)
Ind5	3168	.0126	.0246	.3527	.00392	.00419	.00318	.00113
Other					2.41	3.16^{*}	1.86	2.66
					(.160)	(.016)	(.627)	(.077)
Total	18637	.0087	.0180	.3611	.00298	.00257	.00196	.00122

serve, that all differences are positive with one exception (5vs3 in $PFT - N$ for Industry 3), i.e. from the numerical value we find the non-monotonic behavior. However, only few turn out to be statistically significant. The strongest effect is observed for Industry 2 (Manufacturing), where we find the largest differences which are all strongly significant. On the other hand, for Industries 1 (Consumers) and 4 (Healthcare), the differences are numerically small and far from statistically significant. Interestingly, for these industries the average forecast error $\overline{FE}$ and its associated volatility SFE are small. In contrast, for Industry 2, where the most significant pattern is observed, $\overline{FE}$ and SFE are comparably high. It is well-known that the size and the dispersion of analysts' forecast error are closely related to the information uncertainty (see, e.g. Barron et al., 1998). Thus, we conclude that we find a more pronounced non-monotonous pattern for those industries for which information uncertainty is high, while for industries with lower information uncertainty the pattern is statistically insignificant.

4.3.2 Big versus small companies

Another proxy for information uncertainty which has been used in previous literature (see, e.g., Banz, 1981; Zeghal, 1984; Freeman, 1987) is the size of the company, thereby arguing that larger companies are supposed to have stronger transparency requirements and are covered by more analysts. We therefore identify all observations with respect to the constituents of the Dow Jones Industrial Average (DJI), which covers the 30 largest companies. We analyze the sample including only the DJI companies, as well as our base sample from which we exclude all DJI observations. Results for the ex-DJI sample are on the left side and for only DJI are on the right side of Table 5.[16] The number of observations which refer to DJI companies is 2,071, thereby leaving a size of 16,566 for the ex-DJI sample. Results for the ex-DJI sample are basically similar to the full sample, although we find a slight increase in the statistical difference between category 5 and 4 and 3.[17] The more interesting results refer to the DJI sample (right side), where we observe that the forecast error steadily declines across the performance categories 1 to 5, which suggests that for recommendations with respect to the largest companies, we find the most profitable recommendations to be associated with the smallest forecast error. The statistical post-hoc tests in Panel C confirm this interpretation: By focusing again on category 5, we find significant differences with respect to category 1 and 2, while not being statistically different form category 3 and 4. The use of DJI companies as proxy for companies with smaller information uncertainty may be justified by observing that the mean SE in the DJI sample is 0.0065 with a standard deviation of 0.0161, being significantly smaller then for the ex-DJI sample (mean 0.0089, standard deviation of 0.0182). In analogy to Table 4, panel D shows that for a pairwise comparison of the average forecast errors in

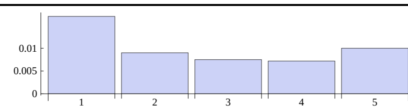
[16] For space reasons, we report only results for the equal interval categories; results for quintiles are qualitatively similar.

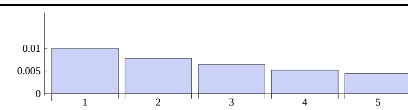
[17] The statistic for 5vs4 increases (in absolute terms) from -3.45 to -3.70; the statistic for 5vs3 increases from -4.05 to -4.30.

Table 5: Sample S&P500ex–DJI vs. DJI

The left side of this table shows results for the full sample exclusive DJI companies for 5 performance categories classified according to equal intervals ($PFT - D$ S&P5exDJI), while the right side shows results for DJI companies ($PFT - Q$ DJI). Panel A reports estimates for the mean forecast error ($\overline{FE}$) and the standard deviation (SFE) together with number of observations (N). Panel B reports results of the Kruskal-Wallis one-way ANOVA to test for equality of distribution. To further analyze between which categories differences are significant, Panel C reports post-hoc tests, where pairwise multiple comparisons are done on the basis of the procedure of Dunn (1964) to adjust for familywise type I error. p-values are below statistics in parantheses. Finally, panel D reports standard t-test results for a comparison of category 5 against category 2, 3, and 4. F reports the Levene-test-statistic for equal variance, while T is the corresponding test-statistic (df reports degree of freedoms).

A. Estimates





	$PFT - D$ S&PexDJI					$PFT - D$ DJI				
	1 (low)	2	3	4	5 (high)	1 (low)	2	3	4	5 (high)
$\overline{FE}$	0.0170	0.0090	0.0075	0.0072	0.010	0.010	0.0078	0.0064	0.0052	0.0045
SFE	0.0291	0.0171	0.0159	0.0141	0.0208	0.0198	0.0191	0.0166	0.0114	0.0095
N	1702	3098	6157	4108	1501	135	403	953	473	107

B. Kruskal-Wallis

	N	df	F	p	N	df	F	p
$\overline{FE}$	16566	4	344.2^{**}	0.000	2071	4	37.8^{**}	0.000

C. Post-hoc tests

	1	2	3	4	5	1	2	3	4	5
1		10.11^{**}	17.26^{**}	15.97^{**}	9.85^{**}		2.59	4.29^{**}	4.69^{**}	5.22^{**}
		(.000)	(.000)	(.000)	(.000)		(.095)	(.000)	(.000)	(.000)
2			7.61^{**}	6.53^{**}	-1.39			2.30	2.96^{*}	3.84^{**}
			(.000)	(.000)	(1.0)			(.214)	(.031)	(.001)
3				-.60	-4.30^{**}				1.13	2.76
				(1.0)	(.000)				(1.0)	(.057)
4					-3.70^{**}					2.03
					(.002)					(.417)

D. t-test against 5

	F	p	df	T	p	F	p	df	T	p
2	20.1	.000	2506	-2.08^{*}	.038	4.97	.026	347	2.50^{*}	.013
3	66.4	.000	2082	-4.89^{**}	.000	2.06	.151	1058	1.17	.241
4	97.3	.000	2025	-5.45^{**}	.000	0.09	.762	578	.58	.561

category 5 (highest profitability) with those 4 and 3 we find a significant positive difference for the ex-DJI sample, while no significant difference is found for the DJI sample.

4.3.3 Standard deviation of the forecast errors

Arguably, an even cleaner study of the impact of information uncertainty may be done by directly sorting the full sample on the basis of an uncertainty proxy, forming quintiles, and comparing them. Therefore, we calculate the volatility of forecast errors on the level of individual companies.[18] We sort on company-level forecast uncertainty, and classify into quintiles labeled SFE_j1 to SFE_j5. We report results for the lowest $SD1$ (highest $SD5$) information uncertainty quintile in the left (right) side of Table 6. First, we observe that the sorting procedure was highly successful in that the average $\overline{FE}$ within the SFE_j1 sample is 0.0013 while being 0.0349 in the SFE_j5 sample, i.e. 26 times as high. Second, and more importantly, we find the non-monotonous pattern where category 5 (most profitable recommendations) displays the second largest forecast error, only for the sample with high information uncertainty (SFE_j5; right side of the Table). For the sample with low information uncertainty SFE_j1, we observe a steady decline in the forecast error across the five performance categories. As before, panels B to D confirm this interpretation statistically. Thus, we are led to conclude that the evidence so far supports the claim that the occurrence of a non-monotonous relationship between forecast accuracy and recommendation profitability is a robust phenomenon which tends to occur in particular under conditions of high information uncertainty.

4.3.4 Regression approach

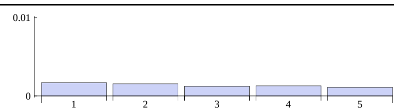
In order to test if our finding is robust to an alternative methodological approach, we use the following empirical procedure. First, we perform univariate regressions on the company-level between forecast error and recommendation profitability. To account for skewness in the predictor variable, we transform both the predictor and response variable with the natural logarithm operator, i.e. we define $LnPFT = \ln(1+PFT)$, and $LnFE = \ln(FE)$.

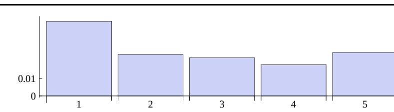
[18] In order to calculate a meaningful standard deviation of FE, we require at least five observations per company. 19 companies do not meet this filter requirement, from which we loose 48 observations, leaving us with 18589 observations on 375 companies.

Table 6: High vs. low information asymmetry

The left side of this table shows results for the 20% of companies with the lowest information uncertainty (SFE_j1), while the right side shows results for the 20% of companies with the highest information uncertainty (SFE_j5). Performance categories are classified on equal intervals ($PFT-D$). Panel A reports estimates for the mean forecast error ($\overline{FE}$) and the standard deviation (SFE) together with number of observations (N). Panel B reports results of the Kruskal-Wallis one-way ANOVA to test for equality of distribution. To further analyze between which categories differences are significant, Panel C reports Post-hoc tests, where pairwise multiple comparisons are done on the basis of the procedure of Dunn (1964) to adjust for familywise type I error. p-values are below statistics in parantheses. Finally, panel D reports standard t-test results for a comparison of category 5 against category 2, 3, and 4. F reports the Levene-test-statistic for equal variance, while T is the corresponding test-statistic (df reports degree of freedoms).

A. Estimates





	$PFT-D\ SFE_j1$					$PFT-D\ SFE_j5$				
	1 (low)	2	3	4	5 (high)	1 (low)	2	3	4	5 (high)
$\overline{FE}$	0.0017	0.00155	0.00124	0.00129	0.0011	0.043	0.024	0.022	0.018	0.025
SFE	0.00158	0.00146	0.00131	0.00129	0.00135	0.045	0.032	0.033	0.027	0.035
N	233	737	1712	946	234	457	647	1135	756	371
B. Kruskal-Wallis										
	N	df	F	p		N	df	F	p	
SE	3871	4	66.5**	0.000		3366	4	162.1**	0.000	
C. Post-hoc tests										
	1	2	3	4	5	1	2	3	4	5
1		2.05	5.76**	4.84**	5.69**		7.41**	11.05**	11.92**	7.25**
		(.397)	(.000)	(.000)	(.000)		(.000)	(.000)	(.000)	(.000)
2			5.62**	4.07**	4.97**			3.23*	4.73*	-.829
			(.000)	(.000)	(.000)			(.012)	(.000)	(1.0)
3				-1.17	1.75				2.01	-1.76
				(1.0)	(.791)				(0.449)	(.777)
4					2.33					-3.14*
					(.196)					(.016)
D. t-test against 5										
	F	p	df	T	p	F	p	df	T	p
2	2.19	.193	978	3.56**	.000	2.43	.119	1016	-.53	.592
3	0.22	.633	1953	.806	.42	3.57	.059	1504	-1.59	.111
4	0.007	.932	1187	1.31	.18	25.6	.000	594	-3.73**	.000

We then estimate the regression model: $LnPFT_j = \alpha_j + \beta_j LnFE_j + \epsilon$ for each individual company j in our sample. We use two regression methods: First, we estimate via least squares with Newey-West standard errors which address autocorrelation and heteroskedasticity in error terms. Second, we estimate via robust regression (more precisely using the MM-estimator) in order to address the potential impact of extreme outliers and leverage points. As mentioned above, after applying filter rules the sample consists of 375 companies, thus, we obtain 375 different β_j coefficients. We only retain coefficients which are significant at the 10%, 5%, or 1% level respectively. On average, β_j coefficients are negative and suggest the negative relationship between forecast errors and recommendation profitability on the company-level. In a second step, we regress β_j coefficients against the proxy for company-level information uncertainty, i.e. the standard deviation SFE_j of forecast errors (per company). Thus, we estimate the regression model $\beta_j^{x\%} = c + b\,SFE_j + \epsilon$, again, both with least squares (Newey-West standard errors) and robust regression (MM-estimator) for the three samples corresponding to the three significance levels $\beta^{1\%}$, $\beta^{5\%}$, and $\beta^{10\%}$. Results of this procedure are summarized in Table 7. The left (right) side of the table regresses on β_j coefficients obtained from least square (robust regression) estimation, while panel A (B) estimates the regression model with LS (robust) regression. The three rows in panel A and B report the regression results for the different significance levels. The important finding is that all coefficients b turn out to be positive, 9 out of 12 are significant at the 5%-level, and 11 out of 12 are significant at the 10%-level. Therefore, we obtain robust evidence that the coefficient between forecast error and profitability (i.e. β_j) gets smaller (in absolute terms) as the proxy for information uncertainty is increasing. This finding supports the view that the relationship between accuracy and profitability is stronger for situations of low information uncertainty, but gets blurred with higher levels of information uncertainty.

4.4 Correlation as explanatory variable

Beside information uncertainty, we hypothesize that the average correlation of recommendations with the market may be an explanatory variable for the observed non-monotonous pattern between accuracy and profitability. By splitting the full sample according to industries, we mentioned that the most pronounced pattern occurred in the Manufacturing industry for which information uncertainty was high, and correlation was below average. In

Table 7: Two-step regression

This table is based on a two-step regression. In a first step, for each individual company j we estimate $LnPFT_j = \alpha_j + \beta_j LnFE_j + \epsilon$ via least squares and robust regression. In a second step we regress those coefficients which are significant at a x%-level against the standard deviation of forecast errors per company SFE_j. Panel A shows results for the OLS regression, panel B for the robust, respectively.

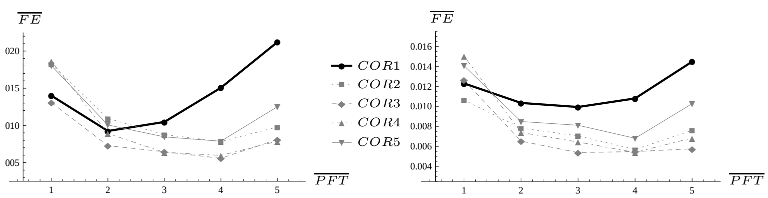
	first regression – OLS β_j					*first regression – robust β_j*				
A. OLS – $\beta_j^{x\%} = c + b\,SFE_j + \epsilon$										
	N	b	stat	p	R^2	N	b	stat	p	R^2
$\beta^{10\%}$	104	0.074*	2.54	0.0125	0.079	98	0.097**	2.79	0.006	0.143
$\beta^{5\%}$	81	0.085*	2.47	0.0153	0.099	71	0.126*	2.61	0.011	0.178
$\beta^{1\%}$	38	0.137*	2.34	0.024	0.245	40	0.150*	2.36	0.023	0.303
B. Robust regression – $\beta_j^{x\%} = c + b\,SFE_j + \epsilon$										
	N	b	stat	p	R^2	N	b	stat	p	R^2
$\beta^{10\%}$	104	0.014	1.18	0.237	0.018	98	0.020*	2.35	0.018	0.067
$\beta^{5\%}$	81	0.025	1.86	0.062	0.06	71	0.016	1.69	0.089	0.047
$\beta^{1\%}$	38	0.041*	2.24	0.024	0.141	40	0.037**	3.23	0.001	0.243

this section, we further analyze the impact of correlation by sorting our full sample according to this variable and classify into quintiles, which we label as $COR1$ (lowest) to $COR5$ (highest correlation). Within each correlation quintile, we form as usual performance categories according to equal intervals or quintiles. We report results in Table 8. In panel A, we present values of $\overline{FE}$ for different correlation and performance quintiles within one graph. The solid, thick line represents results for the $COR1$, i.e. the lowest correlation quintile across the performance categories. Corresponding numerical values are in panel B. Panel C reports post-hoc tests for $COR1$. In Panel D, we perform post-hoc tests *within* a specific performance category but *across* correlation quintiles to test if they differ significantly. Two observations from this analysis are striking. First, within the low correlation quintile $COR1$, the most successful recommendations display by far the largest average forecast error. Second, the pattern for the low correlation quintile $COR1$ is remarkably different from the other correlation quintiles

Table 8: Correlation as explanatory variable

This table reports results for the full sample $N = 18637$. According to the correlation between the returns of the recommendations with the market , we assign all observations to 5 equal-sized groups: from $COR1$ (lowest) to $COR5$ (highest) correlation. Within each correlation quintile we form performance categories according to equal intervals ($PFT-D$) and equal quantiles ($PFT-N$). Panel A shows the corresponding results. For the category $COR1$, Panel B shows the mean forecast error ($\overline{FE}$), the standard deviation (SFE) together with number of observations (N). Panel C reports post-hoc tests for the category $COR1$, where pairwise multiple comparisons between the performance groups are done on the basis of the procedure of Dunn (1964) to adjust for familywise type I error. p-values are below t-statistics in parentheses. Panel D tests whether the trading profitability of $COR1$ in the different performance groups differs from that of the other correlation categories.

	$PFT-D$					$PFT-N$				
A. $\overline{FE}$										



B. Estimates for $COR1$

	1 (low)	2	3	4	5 (high)	1 (low)	2	3	4	5 (high)
$\overline{FE}$	0.014	0.0092	0.0104	0.0150	0.0212	0.0122	0.0103	0.0099	0.0107	0.0145
SFE	0.0216	0.0165	0.0219	0.0265	0.0328	0.0199	0.0201	0.0215	0.0195	0.0252
N	321	715	998	395	146	990	925	704	568	541

C. Post-hoc tests for $COR1$

	1	2	3	4	5	1	2	3	4	5
1		3.98**	4.58**	-0.77	-2.56		2.58	3.84**	0.89	-4.34**
		(.001)	(.000)	(1.0)	(.102)		(.099)	(.001)	(1.0)	(.000)
2			-0.53	-3.34**	-5.77**			1.51	-1.24	-6.33**
			(1.0)	(.008)	(.000)			(1.0)	(1.0)	(.000)
3				-3.96**	-6.21**				-2.44	-7.17**
				(.001)	(.000)				(.145)	(.000)
4					-3.25*					-4.57**
					(.011)					(.000)

D. Post-hoc tests within PFT categories

$COR2$	-0.39	-1.03	2.23	4.81**	4.25**	3.98**	3.94**	7.36**	6.93**	6.82**
	(1.0)	(1.0)	(0.253)	(.000)	(.000)	(0.001)	(.001)	(.000)	(.000)	(.000)
$COR3$	3.30*	4.62**	8.90**	9.83**	9.06**	0.66	6.26**	9.18**	9.49**	9.77**
	(.011)	(.000)	(.000)	(.000)	(.000)	(1.0)	(.000)	(.000)	(.000)	(.000)
$COR4$	-0.17	3.11*	9.05**	9.77**	8.14**	0.09	3.17*	8.38**	7.41**	7.10**
	(1.0)	(0.019)	(0.000)	(0.000)	(.000)	(1.0)	(0.015)	(.000)	(.000)	(.000)
$COR5$	-1.37	0.43	2.52	3.41**	2.14	-0.7	1.59	4.39**	3.80**	0.29
	(1.0)	(1.0)	(0.117)	(0.007)	(.320)	(1.0)	(1.0)	(.000)	(.001)	(1.0)

(displayed in gray). We elaborate on both observations in Panel C and D from the statistical point of view. The post-hoc tests within $COR1$ show that performance category 5 (highest) is strongly significantly different from most other categories in the performance quintiles (right side), while due to smaller sample size we loose some statistical significance when testing with equal interval distance categories (left side). However, overall the statistical tests support the claim that the displayed pattern is robust. Panel D tests if the low correlation category is different from the other quintiles. We find that within performance classes 1 and 2, only few pair-wise comparisons yield significant differences, while within performance classes 4 and 5 only few pair-wise comparisons do *not* yield significant differences. Thus, within the low performance categories, we see no differences in $\overline{FE}$ across correlation categories, while in high performance categories, we observe significant differences. Taken together, the findings suggest the interpretation that successful recommendation trade-off forecast accuracy against correlation. As long as recommendations are weakly related with the market, a high forecast error seems not to be detrimental to the recommendation profitability.

4.5 Multivariate regression approach

The previous sections provide support that the non-monotonous pattern between forecast accuracy and recommendation profitability hinges on information asymmetry (proxied by forecast error volatility) as well as on correlation of the trading strategy with the market returns. In this section, we provide further evidence for these findings by performing multivariate regressions on the analyst level. As in Table 7, we run least squares regression with Newey-West standard errors and robust regressions (MM-estimation) in order to account for outlier impact. We use the volatility of forecast errors on the individual analysts' level in order to gauge the impact of information uncertainty in the regression framework, i.e. we define the variable SFE_i to measure the standard deviation of forecast errors of each analyst in our sample.[19] As our (filtered) sample includes 1679 analysts, the regression in this section has 1679 observations. As further regressors we use the usual forecast error, $\overline{FE}_i$ which is the mean forecast error for individual analysts.

[19] Note that SFE_i is different from our sorting variable SFE_j in the previous section which is the standard deviation of forecast errors for individual companies.

As an alternative measure for the standard deviation, we use the skewness within the forecast errors of analysts. And finally, we use the average covariance of the analysts' recommendations with the market, as proxied by $\overline{COV}_i$. Results are summarized in Table 9. In the first four columns we

Table 9: Multiple regressions

This table reports aggregated results for the single analysts, i.e. $N = 1679$. $\overline{FE}_i$ is the mean forecast error of each analyst i, SFE_i ($SKEW_i$) the standard deviation (skewness) in the forecast errors, and $\overline{COV}_i$ the average covariance of the analyst's recommendation with the market. The dependent variable is the average profitability of each analyst $\overline{PFT}_i$. Panel A reports results for the OLS regression, while Panel B uses robust regression.

A. OLS – Newey-West – Dependent variable $\overline{PFT}_i$

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
c	0.012**	0.012**	0.024**	0.021**	0.031**	0.034**	0.033**	0.045**
	(.000)	(.000)	(.000)	(.000)	(.006)	(.000)	(.000)	(.000)
$\overline{FE}_i$	-1.56**				-1.53**	-1.49**		-1.47**
	(.000)				(.000)	(.000)		(.000)
SFE_i		-1.19**					-1.11**	
		(.000)					(.000)	
$SKEW_i$			-0.010*		-.008			-0.005
			(.0325)		(.104)			(.295)
$\overline{COV}_i$				-166.9**		-163.4**	-161.6**	-162.3**
				(.000)		(.000)	(.000)	(.000)
R^2	.022	.024	.0017	.045	.022	.066	.067	.066

B. Robust regression – MM-estimation – Dependent variable $\overline{PFT}_i$

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
c	0.011**	0.011**	0.017	0.021**	0.023	0.031**	0.031**	0.038**
	(.000)	(.000)	(.154)	(.000)	(.055)	(.000)	(.000)	(.001)
$\overline{FE}_i$	-1.17**				-1.15**	-1.19**		-1.18**
	(.000)				(.000)	(.000)		(.000)
SFE_i		-0.93**					-0.92**	
		(.000)					(.000)	
$SKEW$			-0.006		-0.005			-0.003
			(.1912)		(.313)			(.528)
$\overline{COV}_i$				-161.0**		-161.2**	-160.3**	-160.6**
				(.000)		(.000)	(.000)	(.000)
R^2	.017	.021	.0013	.059	.018	.077	.080	.077

test the univariate model, while columns 5 to 8 test combinations. Note that estimating cross-correlations between the variables reveals that $\overline{FE}_i$ and SFE_i are substantially correlated, so that due to multicollinearity a multivariate regression with both regressors does not yield meaningful coefficient estimates and is therefore omitted. Results are basically robust to the estimation method used. In line with previous results, we find a negative

coefficient on $\overline{FE}_i$ as well as on SFE_i, which is to be expected due to their positive correlation. We find that skewness in analysts forecast error does not contribute to the explanation of performance. Most importantly, we find covariance to be strongly significant with a negative sign both in the univariate as well as in the multivariate setting, where the impact of covariance remains basically unchanged after including the remaining regressors. Thus, in line with what predicted by the model of Lawrenz and Weissensteiner (2012), high covariance is associated with lower performance. Furthermore, although R^2 are generally modest, covariance yields the highest R^2 in the univariate setting.
Thus, from the multivariate regression approach on the analyst level we conclude that the magnitude of the forecast error, as well as their volatility has a negative impact on performance, but so does covariance. If a large forecast error is only weakly correlated, the two effects can offset each other.

5 Conclusion

We investigate the empirical relationship between forecast precision and trading profitability for the constituents of the S&P500. We use earnings per share announcements of the single analysts as proxy for the forecast precision, and use trading recommendations to calculate the excess return over the Libor rate over a holding period of six months. In line with previous studies in this field, we show that recommendations of analysts with the highest forecast precision outperform those with the lowest forecast precision.

As the main contribution of this paper, we show that the empirical relationship between forecast precision and trading profitability need not to be monotonic. Therefore, in a first step we sort our data on the basis of performance and investigate the characteristics of the different groups. In contrast to intuition, we show that the second highest forecast errors occur in the category of the most profitable recommendations. To investigate this non-monotonic relationship we sort our sample according to the 5-industry specification proposed by Kenneth French. We analyze the difference in forecast precision between the group with the highest trading returns and the ones with the second-highest and with intermediate trading returns. The highest differences can be observed for Manufacturing and the remaining category "Other", where absolute forecast errors and the standard deviation is relatively high. Thus, we conclude that the non-monotonic relationship is pronounced in case of high information uncertainty. Furthermore, we use company size as proxy for information uncertainty and compare results for the companies of the Dow Jones Industrial Index with the rest of our sample. In line with our first interpretation, we find the non-monotonic relationship for the smaller companies. We show that we get qualitatively the same results when sorting according to the standard deviation in the forecast errors. A two-step regression support the view that the link between accuracy and profitability is stronger in case of low information uncertainty.

As suggested by previous theoretical papers, we use the covariance of the trading returns with the market as proxy for the covariance in the forecast errors of the single analysts. We show that covariance is an important explanatory variable for the relationship between forecast precision and trading profitability, and that the non-monotonic relationship seems to be most pronounced for companies with a low correlation.

References

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, 9(1):3–18.

Barber, B., Lehavy, R., McNichols, M., and Trueman, B. (2001). Can investors profit from the prophets? security analyst recommendations and stock returns. *Journal of Finance*, 56(2):531–563.

Barron, O. E., Kim, O., Lim, S. C., and Stevens, D. E. (1998). Using anaysts' forecasts to measure properties of analysts' information environment. *The Accounting Review*, 73(4):421–433.

Basu, S. and Markov, S. (2004). Loss function assumptions in rational expectations tests on financial analysts' earnings forecasts. *Journal of Accounting and Economics*, 38(0):171 – 203.

Bhushan, R. (1989). Firm characteristics and analyst following. *Journal of Accounting and Economics*, 11(2–3):255–274.

Bonini, S., Zanetti, L., Bianchini, R., and Salvi, A. (2010). Target price accuracy in equity research. *Journal of Business Finance & Accounting*, 37(9/10):1177–1217.

Bradshaw, M. T. (2004). How do analysts use their earnings forecasts in generating stock recommendations? *The Accounting Review*, 79(1):25–50.

Brav, A. and Lehavy, R. (2003). An empirical analysis of analysts' target prices: Short-term informativeness and long-term dynamics. *Journal of Finance*, 58(5):1933–1967.

Brown, J. and Pfeiffer, J. (2008). Do investors under-react to information in analysts' earnings forecasts? *Journal of Business Finance & Accounting*, 35(7/8):889–911.

Cen, L., Hilary, G., and Wei, K. C. J. (2013). The role of anchoring bias in the equity market: Evidence from analysts' earnings forecasts and stock returns. *Journal of Financial And Quantitative Analysis*, 48(1):47–76.

Clarke, J. and Subramanian, A. (2006). Dynamic forecasting behavior by analysts: Theory and evidence. *Journal of Financial Economics*, 80:81–113.

Clement, M. B. (1999). Analyst forecast accuracy: Do ability, resources, and portfolio complexity matter? *Journal of Accounting and Economics*, 27(3):285–303.

Clement, M. B. and Tse, S. Y. (2005). Financial analyst characteristics and herding behavior in forecasting. *Journal of Finance*, LX(1):307–341.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252.

Easley, D., O'Hara, M., and Paperman, J. (1998). Financial analysts and information-based trade. *Journal of Financial Markets*, 1(2):175–201.

Ertimur, Y., Sunder, J., and Sunder, S. V. (2007). Measure for measure: The relation between forecast accuracy and recommendation profitability of analysts. *Journal of Accounting Research*, 45(3):567–606.

Fischer, P. E. and Verrecchia, R. E. (1998). Correlated forecast errors. *Journal of Accounting Research*, 36(1):91–110.

Freeman, R. N. (1987). The association between accounting earnings and security returns for large and small firms. *Journal of Accounting and Economics*, 9(2):195–228.

Gleason, C. A., Bruce Johnson, W., and Li, H. (2013). Valuation model use and the price target performance of sell-side equity analysts. *Contemporary Accounting Research*, 30(1):80–115.

Gleason, C. A. and Lee, C. M. (2003). Analyst forecast revisions and market price discovery. *The Accounting Review*, 78(1):193–225.

Grossman, S. J. and Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, 70(3):393–408.

Hall, J. L. and Tacon, P. B. (2010). Forecast accuracy and stock recommendations. *Journal of Contemporary Accounting and Economics*, 6(1):18–33.

Hong, H. and Kubik, J. D. (2003). Analyzing the analysts: Career concerns and biased earnings forecasts. *Journal of Finance*, 58(1):313–351.

Huang, J., Mian, M. G., and Sankaraguruswamy, S. (2009). The value of combining the information content of analyst recommendations and target prices. *Journal of Financial Markets*, 12(4):754–777.

Huber, J. (2007). 'J'-shaped returns to timing advantage in access to information - experimental evidence and a tentative explanation. *Journal of Economic Dynamics and Control*, 31(8):2536 – 2572.

Huber, J., Kirchler, M., and Sutter, M. (2008). Is more information always better?: Experimental financial markets with cumulative information. *Journal of Economic Behavior & Organization*, 65(1):86 – 104.

Hule, R. and Lawrenz, J. (2008). The value of information. Some clarifications and some new results for the *Schredelseker-Game*. In Hanke, M. and Huber, J., editors, *Information, Interaction, and (In)efficiencies in financial markets*. Linde, Wien.

Jegadeesh, N. and Kim, W. (2006). Value of analyst recommendations: International evidence. *Journal of Financial Markets*, 9(3):274 – 309.

Kim, O., Lim, S. C., and Shaw, K. W. (2001). The innefficiency of the mean analyst forecast as a summary forecast of earings. *Journal of Accounting Research*, 39(2):329–335.

Lamont, O. A. (2002). Macroeconomic forecasts and microeconomic forecasters. *Journal of Economic Behavior & Organization*, 48(3):265–280.

Lawrenz, J. and Weissensteiner, A. (2012). Correlated errors: Why a monotone relationship between forecast precision and trading profitability may not hold. *Journal of Business Finance & Accounting*, 39(5-6):675Ű699.

Loh, R. K. and Mian, G. M. (2006). Do accurate earnings forecasts facilitate superior investment recommendations? *Journal of Financial Economics*, 80:455–483.

Marinelli, C. and Weissensteiner, A. (2013). On the relation between forecast precision and trading profitability of financial analysts. *Working paper*.

Marinovic, I., Ottaviani, M., and Sørensen, P. N. (2013). Forecasters' objectives and strategies. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2a. North-Holland, Amsterdam.

Markov, S. and Tamayo, A. (2006). Predictability in financial analyst forecast errors: Learning or irrationality? *Journal of Accounting Research*, 44(4):725–761.

Martinez, J. V. (2011). Information misweighting and the cross-section of stock recommendations. *Journal of Financial Markets*, 14(4):515 – 539.

Mikhail, M. B., Walther, B. R., and Willis, R. H. (1997). Do security analysts improve their performance with experience? *Journal of Accounting Research*, 35:131–157.

Mikhail, M. B., Walther, B. R., and Willis, R. H. (1999). Does forecast accuracy matter to security analysts. *The Accounting Review*, 74(2):185–200.

Mikhail, M. B., Walther, B. R., and Willis, R. H. (2004). Do security analysts exhibit persistent differences in stock picking ability? *Journal of Financial Economics*, 74(1):67–91.

Ottaviani, M. and Sørensen, P. N. (2006). The strategy of professional forecasting. *Journal of Financial Economics*, 81:441–466.

Ramnath, S., Rock, S., and Shane, P. (2008). The financial analyst forecasting literature: A taxonomy with suggestions for further reserach. *International Journal of Forecasting*, 24:34–75.

Schredelseker, K. (1984). Anlagestrategie und Informationsnutzen am Aktienmarkt. *Zeitschrift für betriebswirtschaftliche Forschung*, 36:44–59.

Schredelseker, K. (2001). Is the Usefulness Approach Useful? Some Reflections on the Utility of Public Information. In McLeay, S. and Riccaboni, A., editors, *Contemporary Issues in Accounting Regulation.*, pages 135–153. Kluwer, Boston.

Simon, A. and Curtis, A. (2011). The use of earnings forecasts in stock recommendations: Are accurate analysts more consistent? *Journal of Business Finance & Accounting*, 38(1-2):119–144.

Sinha, P., Brown, L. D., and Das, S. (1997). A re-examination of financial analysts' differential earnings forecast accuracy. *Contemporary Accounting Research*, 14(1):1–42.

Stickel, S. E. (1992). Reputation and performance among security analysts. *Journal of Finance*, 47(5):1811–1836.

Stickel, S. E. (1995). The anatomy of the performance of buy and sell recommendations. *Financial Analyst Journal*, 51(5):25–39.

Williams, J. (2013). Financial analysts and the false consensus effect. *Journal of Accounting Research*, 51(4):855–907.

Womack, K. L. (1996). Do brokerage analysts' recommendations have investment value? *Journal of Finance*, 51(1):137–167.

Zeghal, D. (1984). Firm size and the information content of financial statements. *The Journal of Financial and Quantitative Analysis*, 19(3):299–310.