Categorization Bias in the Stock Market*

Philipp Krueger, Augustin Landier, and David Thesmar

First Version: January 2012
This Version: April 2012

Abstract

This paper provides evidence of categorization bias in financial markets. Some investors perceive individual firms through the lenses of industries. Such categorical thinking generates mispricing and predictability in stock-returns. We measure the difference in returns between a firm's official SIC industry and its underlying fundamentals by constructing a basket of closely related firms, based on Hoberg and Phillips (2010a,b). We find that stocks comove strongly with their official industry in the short-term and then revert toward their basket of Hoberg and Phillips comparables. Long-short strategies exploiting mispricing due to industry categorization generate statistically significant and economically sizable risk-adjusted excess returns. We also provide evidence that financial analysts are biased by industry categorization: When a firm's official industry is less representative of its fundamentals, analysts sometimes tend to put excessive weight on information related to the official industry and thus make predictable forecast errors.

JEL-Classification: G10, G11, G14, M41
Keywords: Limited attention, behavioral finance, return predictability, analyst forecasts accuracy

*Krueger is at the Geneva Finance Research Institute, Université de Genève. Landier is at the Toulouse School of Economics. Thesmar is at HEC Paris and CEPR. Philipp Krüger, philipp.krueger@unige.ch, Telephone: +41 (0)22 379 85 69; Augustin Landier, augustin.landier@tse-fr.eu, Telephone: +33 (0)5 61 12 86 88; David Thesmar, thesmar@hec.fr, Telephone: +33 (0)1 39 67 94 12. Landier aknowledges support from the Scor Chair at Fondation Jean-Jacques Laffont. Thesmar thanks the HEC Foundation for financial support.

I Introduction

Even if the underlying reality is continuous, economic agents tend to see the world through coarse categories: a student is an A-student or is not, economies are in a recession or aren't, bonds are \investment grade" or \junk", rms belong to manufacturing or services... Categories exist because memory is limited. It would be ine ciently costly for agents to memorize all characteristics of individuals, events or rms: They use categories as \boxes" in which similar objects can be stored and then tend to consider them as similar when making decisions. The use of representative categories biases expectations as soon as agents forget whether an object is strongly or weakly representative of its category: An agent thinking through categories is biased in a predictable manner vis-a-vis a fully rational agent, making use of a wider information set. For instance, imagine that you grade students as A, B or C based on an underlying continuous score and memorize only the nal grade. A student who got an A but was very close to the B threshold is categorized together with students that are on average better than him. Someone who remembers the actual score can thus predict that you tend to overestimate the level of this particular student. There is experimental evidence that such categorization bias is prevalent. For instance, temperature predicted by non-professional subjects jumps at the beginning of each month, even though forecasts are, on average over the month, unbiased. This is because each forecaster overweights the importance of the month in her intuitive statistical model of temperature (\July is hotter than June"). People therefore behave as if categories were better descriptions of reality than they actually are. Following Mullainathan (2002), we call this bias \categorization bias".

Categorization bias implies a combination of under- and overreaction to information (Mullainathan (2002); Hong, Stein, and Yu (2007)). If new information does not trigger a change in category, learning is abnormally slow vis-a-vis the rational Bayesian benchmark. But, as soon as new, though marginal, information forces a change in categories, agents' expectations over-shoot. For instance, as long as it keeps its Michelin star, a restaurant is

perceived as very good, even if the quality of its food has started declining; then, suddenly, it loses its star and falls strongly into disgrace with the gourmet community. From the point of view of an informed observer of the slow decline in quality of the restaurant, this discontinuous reaction is an error: there is under-reaction, followed by over-reaction of the public. Categorization bias seems to be everywhere in economics, with important consequences: In product markets (where it may explain advertisement spending, as in Mullainathan, Shleifer, and Schwartzstein (2008)), in the labor market (where it could lead to ine cient discrimination), in nance (where it could lead to return predictability, as in Mullainathan (2002)), etc. Yet, and surprisingly, the economics literature o ers little evidence of this bias based on eld data.

This paper shows that categorization bias has sizable e ects on asset prices. Financial markets o er a natural testing ground of categorical thinking because of their wealth of data on expectations (analyst forecasts, stock price movements). Our premise is that some market participants tend to mentally group rms according to an industry classi cation (e.g. the SIC code reported in Compustat). However, such classi cations provide only a noisy and sometimes misleading description of what rms actually do (Hoberg and Phillips (2010a)). Investors relying strongly on industrial categorization neglect some public information: For instance they could use more accurate comparables than simply looking at rms in the same SIC category. By using the heuristics that a rm is \like its SIC industry", such investors thus make predictable valuation mistakes, which in turn create mispricing and predictability in stock-returns. Think of a news shock a ecting the \paper and allied products" industry. A fully rational market participant is aware that some members of this industry, like Rock-Tenn - which does paperboard food packaging - are well represented by the \paper" category, while others, like Schweitzer-Mauduit - which produces cigarette paper and therefore depends on cigarette sales - are not. If investors were fully rational, Schweitzer-Mauduit would react little to \paper" shocks, while Rock-Tenn would react more. If, however, investors are subject to categorization bias, Schweitzer-Mauduit is perceived to be more similar to mainstream paper rms

(such as Rock-Tenn) than it actually is. This leads to a testable implication: Firms (like Schweitzer-Mauduit as opposed to Rock-Tenn) should tend to over-react to their o cial industry shocks, and when such overreaction happens, they should subsequently revert toward their fundamental value.

To operationalize our empirical strategy, we rst develop a method to detect whether a stock is subject to categorization bias at a given point in time. To do this, we rely on a measure of rms' industry fundamentals developed recently by Hoberg and Phillips (2010a,b), who compare rms based on the \business description" section of their 10K ling. We use their data to identify a rm's \fundamental" peers, i.e. rms whose business description is highly similar and thus likely to be a good proxy of fundamentals. For each rm, we then calculate the returns of its \o cial" peers, as de ned by the standard SIC classi cation from Compustat, and those of its \fundamental" peers, as de ned by the Hoberg and Phillips (2010a,b) similarity measure. We show that, at high frequencies (e.g. weekly), rms comove strongly with their o cial industry and weakly with their fundamental peers while the opposite pattern holds at lower frequencies. This sharp reversal is in line with the hypothesis that bounded rationality leads some agents to overweight o cial industry shocks in the short-term and that such mistakes are corrected over time. We test this hypothesis further by constructing portfolios based on the divergence between o cial and fundamental industry returns: If investors tend to mix up a rm with its o cial industry, a highly positive di erence between o cial and fundamental returns should be a signal that predicts lower returns for the rm in the future. Going back to our prior example, if last month, Schweitzer-Mauduit's o cial peers (like Rock-Tenn) have outperformed their fundamental peers (like Alliance One, which grows tobacco), SM is likely to be overvalued, and its stock price will decrease subsequently. In a second step, we re ne this signal by focusing on rms whose stock price has closely followed their \o cial industry" in a given month: This is when categorization bias is most likely to be occurring with high intensity. To return to the Schweitzer-Mauduit example, the overvaluation should be more severe if the stock has closely followed

the \paper" industry in the past month. We label such rms \o cial industry followers".

Consistent with the fact that investors put too much weight on the o cial industry classi cation, we nd that \o cial industry followers", whose o cial peers have out-performed (resp. under-performed) fundamental peers, earn large negative (resp. positive) returns over the next month. A long-short strategy based on this insight yields a monthly abnormal return of 1.57 % (more than 18% annual) with a t-stat of 4.1. We also show that return predictability is more pronounced among smaller rms. This is consistent with bounded rationality: If there is a xed or convex cost associated to evaluating an individual rm, more investors are likely to pay that cost for large rms (for which the stakes are higher) than for small rms. Small rms are thus more prone to categorization bias.

To complement the returns evidence, we then look at actual expectations data: We nd that forecast errors of stock analysts follow a pattern consistent with categorization bias. To show this, we adapt the above methodology to earnings forecast data. We compute the di erence between earnings forecasts of \o cial" peers and the forecasts of \fundamental" peers. This forecast discrepancy measures the extent to which a forecast based on the o cial classi cation is upward biased. For rms for which analysts tend to rely heavily on the consensus in the o cial industry, we nd that analyst expectations tend to be upward biased when the forecast discrepancy is positive. This result is consistent with the hypothesis that some analysts are relying too much on comparisons with rms of the o cial industry, which makes their expectations biased when a rm is not well classi ed.

Our paper contributes to several literatures. The behavioral economics literature has produced models of non-Bayesian inference based on categorical thinking. These models rest on the fact that agents assign situations to categories, and assign a probability one to the most likely category. This non-Bayesian feature is what creates the categorization bias, and explains over- and underreaction to information (see Mullainathan (2002); Hong, Stein, and Yu (2007); Mullainathan, Shleifer, and Schwartzstein (2008) embed

categorization bias in a communication model). In this sense, our paper is related to well documented cognitive biases in psychology such as the salience or representativeness heuristics. This literature is, however, largely theoretical (its aim is to build theoretical frameworks able to account for available experimental evidence); our contribution is to provide direct evidence of categorization bias from eld data. The paper also contributes to the large literature on stock habitats (see Barberis, Shleifer, and Wurgler (2005)). This literature suggests that stocks may comove \too much" when they are traded by similar investors (for instance by index funds when they belong to an index). Our paper suggests that categorization bias is a good candidate to generate excess comovement. In a sense, the o cial classi cation tends to generate stock habitat due to investor categorization biases. Last, this paper is related to the literature on investor attention (see Hong, Lim, and Stein (2000)). This literature shows that mispricing may occur as a by-product of investors' inability to quickly process all relevant information (see also Cohen and Frazzini (2008)). In another recent paper, Cohen and Lou (2012) show that investors have trouble assessing the e ect of industry shocks on diversi ed rms, which generates momentum in their returns as information is slowly impounded into prices. Our paper contributes to this literature by isolating the impact of a well de ned psychological bias.

Section II describes the data. Section III describes the results using returns data. Section IV describes our results on analyst forecasts.

II The Data

II.1 Text-based Network Industry Classi cations

The main data used in this study are the Text-based Network Industry Classi cations (TNIC)[1] developed in Hoberg and Phillips (2010a,b). The Hoberg and Phillips (HP) industry classi cations are constructed by parsing product descriptions from 10K forms and forming word vectors for each rm to compute continuous measures of product simi-

[1] http://www.rhsmith.umd.edu/industrydata/index.html

larity for every possible pair of rms. For any two rms i and j in year t, the data provide a real number in the interval of $[0, 1]$ describing how similar the words used by rms i and j are. Unfortunately, the publicly distributed data do not contain this real number, but rather rm links in the form of (gvkey1-gvkey2 combinations) for which the product similarity score exceeds a certain threshold. Thus, for each rm, the data we use give a list of rms that use similar words to describe their products and therefore are likely to operate in the same product market.

II.2 Return Sample

Our rst set of tests requires stock-level returns data. In order to construct it, we start from all publicly traded securities in the CRSP universe between 1995 and 2009 excluding stocks with sharecodes other than 10 or 11. We match this stock level dataset with the Compustat annual les in order to obtain each rm's $official$ SIC code. We chose the historical SIC code from Compustat $(SICH)$ to de ne a rm's o cial industry (alternative choices of industry classi cation are described as robustness checks). Guenther and Rosman (1994); Kahle and Walkling (1996) give some details about the rules used by Compustat to produce rm-level SIC codes based on the rms' 10K lings. Next, we match this CRSP-Compustat merged dataset with the HP text-based network industry classi cations. These classi cations are available between 1996 and 2008. In order to allow for the HP industry classi cations to be known to investors and to eliminate the possibility of look ahead bias, we lag each rm's network industry classi cation by one year. Finally, we obtain rm level analyst coverage from I/B/E/S by counting the number of analysts who have issued at least one scal year end earnings forecast for a given rm throughout a year. In line with the literature, we set the number of analysts to zero whenever a rm included in the CRSP/Compustat sample cannot be matched with I/B/E/S. We end up with a sample of about 48,000 rm-year observations. For each of these rm-year observations, we then obtain monthly stock prices, returns and shares outstanding. This procedure yields a nal sample of 565,266 rm-month observations

between 1997 and 2009. We also consider weekly returns. Our weekly sample has about 2,600,000 firm-week observations.

We define a firm j's "official" industry return $r_{j,official,t}$, as the equally weighted average of the returns during period t of all firms belonging to the same SIC2 category as firm j. We then define $r_{j,fundamental,t}$, which captures the average return of firms who are fundamentally similar to firm j. It is calculated as the equally weighted return of a portfolio consisting of all firms that are linked to firm j in the Hoberg and Phillips sense in that period. In other words, $r_{j,fundamental,t}$ could be interpreted as capturing shocks to economic fundamentals relevant to firm j. By contrast, a firm's official industry portfolio does not necessarily represent firms that are fundamentally linked to each other, but rather firms that are grouped together according to the official industry classification, which can suffer from plain mistakes (another official industry would be more appropriate) or simply coarseness (a lot of firms are put in the same bag). We consider returns at both weekly and monthly frequencies. Table I provides summary statistics for the main return related variables used in this study.

[Table I about here.]

II.3 Analyst Sample

In the second part of the paper, we examine whether analysts are also subject to industry categorization biases when they issue earnings forecasts. To address this question, we first construct a dataset of forecasts at the analyst level. We start from the I/B/E/S Detail History file and restrict ourselves to fiscal year end EPS forecasts for US firms issued during the current fiscal year (Forecast Period Indicator (FPI)=1). We obtain realized earnings from the Actuals files. As pointed out in Diether, Malloy, and Scherbina (2002), we account for stock splits and make the forecast and actual earnings time series comparable. In doing so, we rely on the CRSP cumulative adjustment split factor extracted from the CRSP Daily files, which is the most reliable and accurate way of performing the

split adjustment (see Robinson and Glushkov (2006)). Next, we remove analyst forecasts which seem to be erroneous (e.g. forecasts which are issued after actual earnings have been announced, forecasts for which actual earnings are announced before the fiscal year end, forecasts which have been confirmed before they have been issued for the first time, etc.). We also trim the tails of the EPS forecast distribution at 1% to reduce the impact of statistical outliers.

During the fiscal year, analysts tend to adjust their forecast of year end earnings at several points in time. Hence, each unit of observation is an analyst-firm-date where the forecast of fiscal year end earnings is reported. In each quarter of the year, we keep only the most recently issued or revised analyst-firm observation. Consider for instance an analyst i who, in the same quarter, issues two forecasts for a firm j with fiscal year end in December, the first one on January 29, 1996 and the second one on February 15, 1996. We keep only the most recent forecast, i.e. the forecast issued on February 15. We then organize the forecast data by forecast horizon, which we define as the difference between the date at which a forecast is issued and the fiscal year end date. We dismiss forecasts issued or revised after the fiscal year ends. Given that we keep one single analyst-firm observation per quarter, we end up with a maximum of four forecasts per analyst-firm-year which differ regarding their forecast horizons. Forecasts issued in the first fiscal quarter have a horizon of T = 3 quarters, forecasts issued in the second quarter have a horizon of T = 2 quarters, forecasts issued in the third quarter have a horizon of T = 1 quarter and forecast issued in the last fiscal quarter have a horizon of T = 0 quarters. We denote analyst $i's$ T quarter ahead forecast of firm $j's$ fiscal year end earnings issued in quarter t as $F(T)_{i,j,t}$.

II.3.1 Forecast Errors

We now compute the accuracy associated with analyst $i's$ T quarter ahead forecast of firm $j's$ earnings issued in quarter t as

$$ForecastError(T)_{i,j,t} = (F(T)_{i,j,t} - A_{j,t})/P_{j,t-4},$$

where $A_{j,t}$ denotes the rm's realized earnings per share at scal year end and $P_{j,t-4}$ is the price that prevailed 12 month prior to scal year end.

In our tests, we also use rm-level forecast errors, which are obtained by collapsing analyst- rm level forecast errors with identical horizon T at the rm-quarter level. Formally, we de ne the average bias across rm $j's$ T quarter ahead EPS forecasts in quarter t as

$$ForecastError(T)_{j,t} = \frac{1}{N(T)_{j,t}} \sum_{i=1}^{N(T)_{j;t}} ForecastError(T)_{i,j,t},$$

where $N(T)_{j,t}$ refers to the number of issued or revised T quarter ahead forecasts for rm j in quarter t. For a given forecast horizon T, we thus end up with one yearly observation per rm j.

II.3.2 Consensus Forecasts

Next, we match the stock price that prevailed at the beginning of the quarter in which a forecast is issued and de ne the T horizon forecast to price ratio for analyst $i's$ EPS forecast for rm j in month t as:

$$FP(T)_{i,j,t} = F(T)_{i,j,t}/P_{j,t-1}.$$

II.3.3 Firm-level consensus forecast

We then calculate the average forecast to price ratio at the rm level by averaging across all issued or revised forecasts for rm j in quarter t with a horizon of T quarters, i.e.

$$FP(T)_{j,t} = \frac{1}{N(T)_{j,t}} \sum_{i=1}^{N(T)_{j;t}} FP(T)_{i,j,t}.$$

II.3.4 Industry-level consensus forecasts

For each firm-quarter, we finally calculate forecast to price ratios of the firm's "official" and "fundamental" industries. To do so, we match the forecast data with historical SIC codes from COMPUSTAT, and calculate, for each firm-quarter-horizon triplet, the average forecast to price ratio across all firms belonging to the same SIC2 industry. We will refer to this variable as $FP(T)_{j,official,t}$, the consensus T quarter ahead forecast in firm j's official industry at date t. Analogously, we use the Hoberg and Philipps classification and calculate, for each firm j at date t, the average T quarter ahead forecast to price ratio of all firms that are related to firm j in the Hoberg and Phillips sense. We denote this variable as $FP(T)_{j,fundamental,t}$.

[Table II about here.]

Table II shows summary statistics for the main analyst related variables. Panel A contains variables at the forecast (analyst-firm-year) level. Panel B shows variables at the firm-year level (e.g. industry-level forecast to price ratios). Consistent with the existing literature, the average and median forecast errors at longer horizons are small, but positive. In contrast, bias decreases at shorter forecast horizons (e.g. T = 0). Finally, Panel C shows variables with variation at the analyst-year level: the typical analyst tracks about three SIC2 industries and 11 stocks and has been providing forecasts for about six years. All variables are trimmed at 1% and 99%.

III Stock-price over-reaction to official industry returns

The idea we have in mind is that investors in the stock market overemphasize the representativeness of official industry classifications. If categorical thinking is at work, we thus expect the following: In the short run, stocks exhibit comovement with their official industry peers in excess of what is granted by fundamentals. With time, investors reverse

mistakes induced by such categorical thinking, which then leads prices to revert toward economic fundamentals at longer horizons. Consider, for instance, a firm XYZ for which the official industry is only weakly representative. Such limited representativeness could be due to the firm selling products different from those offered by the typical firm in its industry, or simply by the coarseness of the classification. If some investors put all firms of that official industry in the same box when analyzing information flow, they will create excess comovement between XYZ's returns and its official industry returns. This comovement will subsequently be followed by a reversal toward XYZ's fundamentals. Hence, comovement between imperfectly related firms arises only because investors treat firms that belong to the same official industry category as being highly similar. Thus, our hypothesis is that at high frequencies, firms comove strongly with their official industry and weakly with their fundamental peers while the opposite pattern should hold at lower frequencies.

We first provide a graphical test of our hypothesis using weekly return data. For each stock, we compute cumulative returns over different horizons. Let $r^T_{j,t}$ denote the T week cumulative return of firm j, between time t and $t + T$. Correspondingly, $r^T_{j,fundamental,t}$ and $r^T_{j,official,t}$ denote the cumulative return between week t and week $t + T$ of the firm's official industry and that of the basket of Hoberg-Phillips comparables. We regress a firm's T week cumulative return on the firm's fundamental and the firm's official T week cumulative industry returns. This is done in a pooled regression with double clustered standard errors (time and stock dimension). Formally, we estimate the following equation for horizons T running from 1 to 8 weeks:

$$r^T_{j,t} = a_T + b_T \times r^T_{j,official,t} + c_T \times r^T_{j,fundamental,t} + e^T_{j,t} \quad (1)$$

[Insert Figure 1 here.]

We plot the coefficient estimates b_T and c_T against the return horizon T in figure 1.

The dashed lines represent 95 % con dence intervals. A strikingly clear pattern emerges from the picture: in the short run (one week horizon, T=1), rm level returns are more strongly related to their o cial than to their fundamental industry returns. This is equivalent to saying that at the one week horizon, returns are more sensitive to o cial returns than they are to fundamental returns (as proxied by the returns to a porfolio of rm's HP comparables). By contrast, and as expected, this relationship reverts at longer return horizons: Firm returns become signi cantly more sensitive to their \fundamental" than to their \o cial industry" for horizons of three or more weeks ($T > 2$). The coe cient estimate c_T, i.e. the sensitivity of cumulative stock returns with respect to the fundamental industry return increases monotonically in T, while, simultaneously, the sensitivity with respect to the o cial industry (b_T) decreases monotonically. The gure thus shows that in the short term, stock returns comove $excessively$ with their o cial industries, while in the medium and longer term, they revert toward fundamentals, thus becoming more correlated with HP-peer returns. This second comovement test also validates the use of HP-peers as a proxy for fundamentals: the Hoberg and Philipps classi cation explains stock returns at a one month horizon better than the SIC2, which suggests that it contains more information.

III.1 Portfolio Tests

III.1.1 Basic Test

The graphical evidence presented in the previous section suggests a strong pattern of return predictability based on the idea of industry categorization bias, which we now test by forming portfolios. In this second step, our identi cation strategy focuses on instances in which returns of a rm's $fundamental$ ($r_{j,fundamental,t}$) and $official$ industries ($r_{j,official,t}$) diverge. Categorization bias implies that a strong positive divergence between o cial and fundamental industry shocks (large $r_{j,official,t} - r_{j,fundamental,t}$) should signal strongly negative returns, as the rm's stock price subsequently reverts to fundamentals. Note

that to eliminate any possibility of look ahead bias, we lag the HP industry classi cation by one year when constructing each rm's \fundamental" industry portfolio, and use COMPUSTAT's historical SIC.

We build a simple portfolio strategy based on this insight by sorting stocks according to the industry return di erential ($r_{j,official,t} - r_{j,fundamental,t}$). We start with weekly data. At the beginning of each week, we sort rms according to the return di erential in $t-n$. In total, we consider lags of up to 6 weeks (n = 6). The rst quintile contains stocks for which the return di erential is the most negative: the rm's fundamental industry return strongly exceeds that of its o cial industry. In the fth quintile, o cial industry peers strongly outperform fundamental ones. We expect stocks in the rst quintile portfolio to revert positively to their fundamentally linked rms, while rms in the fth quintile should revert negatively. We restrict all tests to stock for which the price at the beginning of the period exceeds $5. In addition, we require the o cial and the fundamental industries to be populated by at least ve di erent rms.

[Table III about here.]

In table III we report four factor alphas alongside loadings for the Fama and French (1993) and Carhart (1997) factors for the Q1 and the Q5 portfolios. We also show results for a long-short (Q1-Q5) portfolio, which buys rms in the rst and sells rms in the fth quintile of the industry return di erential distribution. All portfolios are equally weighted. In Panel A we sort rms on the industry return di erential at the beginning of the week (t-1). In panel B, we use the industry return di erential at the beginning of the previous week (t-2) and so forth. Consistent with the hypothesis that investors overreact to shocks to a rm's o cial industry, we nd strong evidence that prices revert to their fundamentally linked rms following sizable divergence between o cial industry returns and their fundamentals. The weekly four factor alpha of an equally weighted portfolio which is long in rms with the most negative prior industry return di erential is 28 basis points. With a t-stat of 7.18, this risk adjusted excess

return is highly statistically signi cant. Similarly, rms that have experienced the most positive industry return di erential over the previous week revert negatively. The Q5 portfolio yields a statistically signi cant weekly alpha of -9 basis points (t-stat: -2.33). The long-short portfolio Q1-Q5 yielding 37 basis points is also highly signi cant (t-stat: 5.66). It corresponds to an annualized return of about 19%.

Since we construct portfolios according to the industry return di erential at di erent lags, we are able to examine the frequency at which the reversal occurs. The table reveals that the speed of reversal depends strongly on whether the industry return di erential is positive or negative. Stocks which have been subject to unjusti ed upward price pressure (High fundamental{o cial di erential; Q5 rms) revert fully within a week's time. Beyond that time, the signal fades out: The risk adjusted excess returns are no longer statistically di erent from zero after the rst week. By contrast, the "long" signal of the strategy is much more persistent: "Q1 rms" take much longer to revert completely. The weekly equally weighted alpha decreases somewhat gradually from about 28 basis in the rst, to about 16 basis points in the sixth week. The signal no longer produces signi cant risk adjusted excess returns for the long-short portfolio after three weeks, suggesting that the mispricing dissipates roughly within a month's time. Guided by this observation, we reproduce the portfolio analysis at the monthly level in the appendix (see table A.I) yielding identical conclusions, and we will focus on monthly returns in the remainder of the text.

[Table IV about here.]

III.1.2 Sorting by Size

Bounded rationality theories predict that agents have the choice to \reduce their bias" when the bene t of doing so outweighs the cost (see,for instance, Gabaix (2010)). In the present context, we therefore expect that categorization bias is less pronounced for large caps, as the dollar value of mispricing generated by a similar cognitive mistake is higher for large than small rms. In table IV, we turn to examining risk adjusted excess returns

of value weighted portfolios. Following up on the prior insights regarding the speed of the reversal, we now focus on monthly returns. We first sort all firms according to their market capitalization in the prior month and group them into terciles. Within each size tercile, we then sort firms into five quintile portfolios based on their industry return differential over the previous month. In total, we obtain 15 size-industry differential and three long-short portfolios. Panel A, B and C report the results for all value weighted portfolios consisting of small, medium and large capitalization firms. The results show significant monthly portfolio alpha for long-short portfolios consisting of small (154 basis points; t=4.25) and mid cap firms (119 basis points; t=2.39). By contrast, value weighted portfolios containing large cap firms do not yield significant risk adjusted excess returns. All in all, categorization bias seems to be prevalent in the bottom two terciles of firm size, and rather absent in the top tercile, which is largely consistent with bounded rationality.

III.1.3 Refining the strategy

If categorization bias explains the stock price anomaly we document, the pattern of reversal should be more pronounced when firms have indeed closely followed their "official" industry. This would be the case, for instance, if stock miscategorization fluctuates over time. Sometimes, the stock is properly categorized, and even though fundamental and official peers diverge, the stock follows its fundamental peers and therefore does not revert. Sometimes, the stock is miscategorized, and hence the peer-fundamental divergence signals future reversal.

To test this, we refine our strategy by the means of a double sort. We measure (1) the official–fundamental divergence as above, and (2) the extent to which the return of a stock has been abnormally close to its official industry return in a given month. Our second signal is computed as follows: For each firm-month pair, we define an $Official\ Industry\ Follower$ measure as

$$\rho_{j,t} = \frac{r_{j,t} - r_{j,official,t}}{\sigma_{j,24}(r_{j,t} - r_{j,official,t})},$$

where $r_{j,t}$ is rm j's return in month t, $r_{j,official,t}$ denotes the equally weighted return of the rm's o cial industry and $\sigma_{j,24}(r_{j,t} - r_{j,official,t})$ is the standard deviation of the di erence between rm j's return and the return of the rm's o cial industry calculated for rolling windows of 24 months. Low values of ρ indicate that rm j's stock price change has closely followed the price changes in its o cial industry in month t, while large values indicate divergence. We choose this measure over the rolling correlation between $r_{j,t}$ and $r_{j,official,t}$ since we are ultimately interested in identifying cases in which stocks that typically do not follow their o cial industry-level returns have been abnormally close followers of their o cial industries. Relying on rolling correlations and selecting stocks for which the rolling correlation is high would identify rms that generally tend to follow their o cial industries, while we are interested in identifying cases in which o cial industry following is unusual (low values of ρ).

We now build a portfolio strategy based on a combination of the *Official Industry Follower* measure $\rho_{j,t}$ and the industry return di erential $r_{j,official,t} - r_{j,fundamental,t}$. We rst sort rms into three terciles of ρ over the previous month. The rst tercile marks *Strong Official Industry Follower*, i.e. stocks that have moved unusually strongly in line with their o cial industries over the past month. In contrast, the third tercile of ρ indicates stocks which have not followed their o cial industry *(Weak Official Industry Follower)* over the previous month. Within each tercile of ρ, we then sort rms into quintiles of the previous month's industry return di erential. Thus, the signal we use in order to construct the re ned portfolio strategy is based on the following interaction term $(r_{j,official,t-1} - r_{j,fundamental,t-1}) \times \rho_{j,t-1}$. Figure 2 illustrates the idea of the interacted signal in a simple diagram.

[Figure 2 about here.]

We expect return predictability to be strongest among *Strong Official Industry Follower* stocks. This is because rms which have not followed their o cial industries *(Low Official Industry Follower)* are more likely of having moved in line with their fundamentally linked rms, reducing the scope for return reversal. Hence, the magnitude of the

return reversal should be increasing in the extent to which a stock has moved in line with its official industry peers.

[Table V about here.]

In table V, we report the risk adjusted returns and factor loadings for equally weighted portfolios, which are constructed according to a double sort on $\rho_{j,t-1}$ and $r_{j,official,t-1} - r_{j,fundamental,t-1}$. The results show that return predictability, as evidenced by the risk adjusted excess returns of the long-short (Q1-Q5) portfolio, is monotonically declining in the *Official Industry Follower* measure: the effect is most pronounced among *Strong Official Industry Follower* stocks (158 basis points; t=4.10; see Panel A) and weakest for *Weak Official Industry Follower* stocks (77 basis points; t=1.73; see Panel C). A monotonically decreasing effect is consistent with the view that stock reversal to fundamentals is more pronounced if a stock has unusually closely tracked its official industry in the first place.

[Table VI about here.]

We now restrict the analysis to *Strong Official Industry Follower* stocks (first tercile of $\rho_{j,t-1}$), and form value weighted portfolios at the monthly level. At the beginning of each month, we first sort all *Strong Official Industry Follower* stocks into terciles of market capitalization in the previous period. Then, within a size tercile, we sort on the industry return differential. Comparing value weighted long-short portfolios restricted to *Strong Official Industry* followers with portfolios which are not conditioned on the *Official Industry Follower* measure (see Panels A, B and C of table IV) shows that, indeed, reversal is much more pronounced in both economic and statistical terms for *Strong Industry Follower* stocks. For mid cap stocks, for instance, the monthly alpha of the value weighted long-short portfolio increases from 119 (t=2.39; see Panel B of table IV) to 160 basis points (t=3.05; see Panel B of table VI). This increase is even more pronounced when looking at small capitalization firms. In the appendix (see Table A.II),

we evaluate equally weighted portfolios restricted to *Strong Official Industry Followers* at the weekly frequency. The same dramatic increase of both economic and statistical signi cance for o cial industry movers is evidenced at the higher frequency too: four factor alpha for the long-short portfolio in the rst week increases from 37 (t=5.66; see Panel A of table III) to 48 (t=7.34; see Panel A of table A.II) weekly basis points, when we additionally condition on the *Official Industry Follower* measure over the previous week.

III.2 Regression Tests

As an alternative to the calendar time portfolio approach, we now use cross sectional regression frameworks in order to test our hypothesis. Relying on regression tests allows us to control more thoroughly for other potential cross sectional determinants of stock returns, most notably the rm's lagged return $r_{j,t-1}$, the rm's past cumulative returns $r_{j,t-2,t-12}$, the log of the rm's market capitalization and a rm's book to market equity. We focus on monthly returns in the regression tests.

[Table VII about here.]

In column (1) of table VII, we start by regressing the monthly raw return on the industry return di erential in the previous month. We double cluster standard errors in the month and stock dimension in the spirit of Petersen (2009) and also include month xed e ects in the equation. As expected, the coe cient estimate for $r_{j,official,t-1} - r_{j,fundamental,t-1}$ has a negative sign and is statistically signi cant: when the o cial industry has outperformed the fundamental one, future returns are negative as the stock reverts to its fundamental value. Mirroring the portfolio analysis, we then re ne the analysis in column (2) of table VII by interacting the return di erential ($r_{j,official,t-1} - r_{j,fundamental,t-1}$) with dummy variables indicating *Strong*, *Medium* and *Weak Official Industry Follower* stocks. Consistent with the portfolio tests, the return reversal e ect is most pronounced among *Strong Official Industry Follower* stocks (the

reference category). The marginal effects for the reference category increases both in economic magnitude and statistical significance (t=-3.24) when contrasted with the pooled estimate in column (1). In line with the evidence presented in the portfolio analysis, the effect of the industry return differential decreases monotonically in the three levels of *Official Industry Follower*. Equations (3){(6) implement further controls, but to make results easier to read, we focus on *Strong Official Industry Follower* stocks, for which the mean reversion is the most pronounced. Column (3) shows an economically strong and statistically significant impact of the industry return differential even after controlling for commonly used cross sectional determinants of stock returns. Column (4) interacts the industry return differential with market capitalization and confirms that the categorization bias is weaker for large capitalization firms, which is consistent with bounded rationality as noted earlier. In column (5), we test whether information production, as proxied by analyst coverage, tends to attenuate the bias. We find no evidence in favor of this view.

One potential alternative interpretation of our results could be that return predictability is not due to the categorization bias we have in mind, but rather due to indexing by passive exchange traded funds. Assume for instance that industry-focused ETFs receive flows based on fundamental information about the representative firm of their industry. In this case, such ETFs would exert price pressure on the representative firm but the price would not revert as the flows were motivated by fundamental trading. Stocks that are misclassified in the industry, however, would first comove with the industry portfolio, and then return to their fundamental value. Misclassified stocks would comove too much with their official industry because of ETF flows. The problem would be less one of investor categorization, than one of financial market imperfection due to the low granularity of ETF definitions.

[Figure 3 about here.]

We provide a simple test in column (6). One implication of this alternative interpretation is that the effect should be stronger in recent periods. Figure 3 plots the monthly

dollar volume of the $SPDR\ S\&P\ 500\ (Ticker\ Symbol:\ SPY)$ as a fraction of total monthly dollar volume of the CRSP Universe and shows that growth in ETF trading has picked up remarkably after 2002. We code a dummy variable that marks all firm-month observations after 2002 and interact it with the industry return differential. The interaction effect is not statistically significant, suggesting that our results are not driven by indexing.

Furthermore, in the appendix (see table A.III), we re-estimate the first five specifications of the previous table in a Fama and MacBeth (1973) framework. In calculating the t-statistics, we allow for heteroscedasticity and serial correlation of up to 12 months in the spirit of Ponti (1996). The coefficient estimates resulting from the Fama-Macbeth (FMB) regressions are quite similar in terms of their economic magnitudes and statistical significance is, if anything, stronger in the FMB framework.

III.3 A Placebo Test: Using Pseudo-Fundamental Peers

A potential concern in interpreting our results is that they might be exclusively driven by the returns of the official industry portfolio $r_{j,official,t}$ rather than being due to divergence between the official and the fundamental industry portfolio returns $r_{j,official,t} - r_{j,fundamental,t}$. The observed reversal in the comovement intensity of stocks with their official and fundamental peers (Figure 1) already alleviates this concern somewhat. However, we also address this issue more directly within a portfolio analysis framework by constructing a pseudo Hoberg and Phillips (2010a,b) industry classification by the means of the following algorithm: For each firm j included in the TNIC data in year t, we randomly select l_{jt} different firms from the Compustat universe in year t, where l_{jt} is the number of Hoberg and Phillips links of firm j in year t. We then calculate the equally weighted return to a portfolio consisting of these pseudo Hoberg and Phillips firms and denote the pseudo Hoberg and Phillips portfolio return as $r_{j,pseudo,t}$. For each firm-month observation, the pseudo portfolio contains the same number of firms as the Hoberg and Phillips industry portfolio. Yet, the firms in the pseudo portfolio have no meaningful relation to firm j because they are randomly chosen. We now reproduce the portfolio analysis

for the pseudo return di erential by rst sorting stocks according to whether they track their o cial industries over the previous month (i.e. terciles of $\rho_{j,t-1}$) and then sorting stocks within a given tercile of $\rho_{j,t-1}$ by quintiles of $r_{j,official,t-1} - r_{j,pseudo,t-1}$. Hence, instead of sorting on the industry-return di erential, we now sort stocks into quintiles of $r_{j,official,t} - r_{j,pseudo,t}$. All portfolios are equally weighted.

[Table VIII about here.]

Panels A, B and C of table VIII show risk adjusted excess returns and factor loadings for $Strong,\ Medium$ and $Weak\ Industry\ Follower$ stocks. The pseudo strategy goes long in stocks for which $r_{j,fundamental,t-1} - r_{j,pseudo,t-1}$ is the most negative over the previous month and sells stocks for which the pseudo industry return di erence is most positive. Assuming that $r_{j,pseudo,t-1}$ is just random noise, such a strategy is akin to an inverse industry momentum strategy (buying short term losers and selling short term winners). Consistent with the idea of industry momentum (see Moskowitz and Grinblatt (1999)), we nd that the long-short portfolio Q1{Q5 based on the pseudo return di erential yields statistically signi cant negative four factor alpha. Negative alpha is in stark contrast to the positive risk adjusted returns we document for strategies based on $r_{j,official,t-1} - r_{j,pseudo,t-1}$. In unreported regression tests, we obtain a statistically signi cantly positive coe cient estimate for $r_{j,official,t-1} - r_{j,pseudo,t-1}$, which again is in stark contrast to the negative coe cient for $r_{j,official,t-1} - r_{j,fundamental,t-1}$. This evidence suggests that it is not $r_{j,official,t-1}$, but rather the relationship between the o cial and the fundamental industry returns, i.e. $r_{j,official,t-1} - r_{j,fundamental,t-1}$ that is driving our results.

III.4 Additional Robustness Checks

In untabulated analysis, we de ne a rm's o cial industry at the SIC3 level obtaining similar results. We also obtain qualitatively similar results by using the Global Industry Classi cation Standard (GICS) instead of the Standard Industrial Classi cation. In additional robustness checks, we test whether our results are sensitive to de ning industry

level returns as medians, equally weighted or value weighted averages and find this not to be the case. In a last robustness check, we also find that using DGTW characteristics-adjusted stock returns (see Daniel, Grinblatt, Titman, and Wermers (1997)) leads to similar conclusions.

IV Industry Categorization Bias in Analysts' Forecasts

The aim of this section is to provide evidence of categorization bias directly based on analyst expectations, instead of looking at prices. Here, we take advantage of the wealth of data on realized and expected analyst forecasts. The empirical approach will be similar to the one used in the stock returns section, except that we replace returns by true expectations.

Ideally, in the spirit of Table III, we would want to regress EPS forecast errors on the divergence between "official" and "fundamental" industry forecasts. The idea is that if expectations about "official" peers are, say, better than expectations about "fundamental" peers and *if* analysts suffer from categorization bias, analyst forecasts should be too optimistic: the average firm-level EPS forecast error should be positive. Put differently, divergence between official and fundamental industry forecast predicts systematic biases in expectations, an apparent violation of the rational expectation hypothesis. This naive version of the test is, however, difficult to implement as forecast errors are affected by many observable and unobservable determinants.

This is why we adopt the approach of Table V: we test if forecast errors are *more* affected by official–fundamental industry divergence, *when* analyst expectations tend to track the "official" industry consensus more closely. This conditional approach is akin to a difference-in-difference methodology and identical to the test we run on stock returns in Table V. The main advantage of this approach is that it refines the predictive power of the categorization hypothesis as it allows focusing on instances where (1) the scope for

categorization error is the biggest (official and fundamental industries diverge) and (2) analyst do focus a lot on what is going on in the official industry. In addition, this refined approach allows to better control for unobservable determinants of forecast errors.

First, we define our measure capturing how close forecasts for firm j are at time t to the average forecast prevailing in its the official industry as:

$$\rho(T)_{j,t} = \frac{|FP(T)_{j,t} - FP(T)_{j,official,t}|}{\sigma_t(FP(T)_{j,t} - FP(T)_{j,official,t})}.$$

$FP(T)_{j,t}$ is the firm's consensus forecast to price ratio and $FP(T)_{j,offical,t}$ denotes the average forecast to price ratio prevailing among a firm's official peers. Since we study different forecast horizons, this measure is also indexed by the horizon T. σ_t denotes the cross-sectional standard deviation. $\rho(T)_{j,t}$ measures whether the consensus forecast for firm j is closer to the official industry forecast for firm j than for other firms. In other words, low values of ρ indicate that analysts tend to be close to the official industry level average forecast for firm j, while large values indicate that the consensus forecast for firm j diverges from its official industry-level consensus forecast. In contrast to the returns section, we calculate $\rho(T)_{j,t}$ by normalizing by the cross sectional standard deviation. We do so because the lower frequency of analyst forecasts makes it difficult to use the time series of past forecasts as we did in our stock returns tests. Thus, $\rho(T)_{j,t}$ closely mirrors the measure we adopted in our stock returns tests of Table V.

We then implement our test by running the following regression:

$$\begin{aligned} ForecastError(T)_{j,t} = {} & \alpha + \beta(FP(T)_{j,offical,t} - FP(T)_{j,fundamental,t}) + \gamma\rho(T)_{j,t} \\ & + \delta\rho(T)_{j,t} \times (FP(T)_{j,offical,t} - FP(T)_{j,fundamental,t}) + \lambda' X_{j,t} + \theta_t + \epsilon_{j,t}. \end{aligned}$$

where $ForecastError(T)_{j,t}$ is the average T horizon firm-level forecast-error, $FP(T)_{j,official,t} - FP(T)_{j,fundamental,t}$ is the divergence between forecasts of official and fundamental peers, and $\rho(T)_{j,t}$ is our measure of analysts' official industry focus. We run these regressions separately for different forecast horizons $T = 0, 1, 2, 3$. In the specifications, we also

control for a number of firm level characteristics (e.g. book-to-market, size, earnings volatility, etc.) that have been used in the literature to explain forecasts errors (see Hong and Kacperczyk (2010)) and include year fixed effects. The appendix contains a complete list of the firm-level controls used.

In the above specification, our hypothesis is that $\delta < 0$. If analysts have the tendency to closely track the official industry level forecast for a firm (small $\rho(T)_{jt}$) and there is strong positive divergence between official and fundamental earnings prospects ($FP(T)_{j,offical,t} - FP(T)_{j,fundamental,t} > 0$), analysts will tend to be overly optimistic.

[Table IX about here.]

We report the results for several specifications and all four forecast horizons in table IX. In columns (1){(4), we examine long horizon forecasts, i.e. forecasts issued in the first fiscal quarter (horizon T = 3), while columns (5), (6) and (7) consider bias associated with shorter horizon forecasts (T = 2, 1, 0 quarters respectively). In columns (1) and (2), we regress the three quarter ahead firm-level forecast error on the official-fundamental divergence. As noted before, the sign of this coefficient is difficult to interpret because of unobservable characteristics influencing forecast errors: whether we control for observable characteristics or not, it is statistically indistinguishable from zero. In columns (3)-(7), we interact the fundamental-official divergence with our firm-level measure of analysts' official industry focus $\rho(T)_{jt}$. Column (3) looks at a forecast horizon of 3 quarters (T = 3): the coefficient estimate for the interaction term is significantly negative. This is consistent with what we expect: whenever the fundamental forecast is lower than the official one ($FP(T)_{j,official,t} - FP(T)_{j,fundamental,t} > 0$), and analysts focus a lot on official peers ($\rho(T)_{j,t}$ small), then forecast errors are very positive. In these instances, analysts are too optimistic and their bias can be systematically predicted. In column (4) we add firm level control variables to the equation which leaves our conclusions unchanged. In columns (5) to (7) we repeat this exercise for shorter forecast horizons (T = 2, 1, 0). The interaction term is significant for a forecast horizon of two quarters and becomes insignificant at

shorter horizons. The observation that categorization bias decays at shorter horizons is consistent with a bounded rationality argument: at shorter horizons, public information about future EPS is more abundant (in particular because of quarterly accounts). Hence, it is less costly for analyst to debias their expectations, so they become more rational.

We now study whether the tendency for analysts to be biased is stronger for speci c kinds of analysts. Again, we try to isolate at the rm-level cases in which anchoring is particularly problematic (high o cial-fundamental divergence), and ask which analysts have, in these instances, a tendency to track the o cial industry in their forecasts. It should be noted that anchoring on a rm's o cial industry consensus is not bad per se. This is because it might be a good strategy to use the o cial industry average as an anchor whenever the rm in question is very similar to its o cial industry peers. By contrast, o cial industry emphasis is problematic at times when a rm's o cial and true industries di er substantially.

To implement this test, we rst need to calculate the extent to which analyst $i's$ forecast for rm $j's$ year end earnings is close to the o cial industry by adapting our ρ measure to individual analyst forecasts:

$$\rho(T)_{i,j,t} = \frac{|FP(T)_{i,j,t} - FP(T)_{j,official,t}|}{\sigma_{jt}(FP(T)_{i,j,t} - FP(T)_{j,official,t})}.$$

In calculating $\rho(T)$ at the analyst level, we now choose to normalize by the cross-sectional standard deviation at the rm-year level (σ_{jt}). We do so because we are ultimately interested in capturing whether an analyst i compared to other analysts forecasting earnings for the same rm j tends to anchor more strongly on the o cial industry.

We then regress $\rho(T)_{i,j,t}$ on an interaction term between the absolute fundamental-o cial divergence $|FP(T)_{j,official,t} - FP(T)_{j,fundamental,t}|$ and observable analyst characteristics. We expect more rational analysts to deviate more from the o cial industry consensus, when such deviations are justi ed, i.e. the o cial category provides a bad approximation (high divergence $|FP(T)_{j,official,t} - FP(T)_{j,fundamental,t}|$). In other words, more

rational analysts should have higher $\rho(T)_{i,j,t}$ when $|FP(T)_{j,official,t} - FP(T)_{j,fundamental,t}|$ is relatively high. In all regressions, in addition to the $industry\ divergence \times analyst\ characteristics$ interaction term, we include as controls several rm and other analyst characteristics (e.g. number of stocks or industries covered by the analyst, etc.) as well as analyst and year xed e ects.

[Table X about here.]

Taken together, our results suggest that more informed analysts tend to be more rational (i.e. deviate from the consensus) when such deviations are required (high fundamental-o cial divergence). We report results in Table X, and focus on long horizon forecasts (T=3), where biases seem to be the strongest according to results from Table IX. We rst regress $\rho(3)_{i,j,t}$ on the industry di erence and year xed e ects. The coe cient estimate is signi cantly positive, suggesting that on average analysts deviate from the o cial consensus (see column (1)). This suggests that, on average, when o cial and fundamental signals diverge, analyst forecasts tend to track the o cial industry less closely ($\rho(3)_{i,j,t}$ is bigger). On average, analysts trust o cial signals less when they are less relevant. As we have seen in the previous Table, this correction is not large enough to prevent the consensus from being biased, but analysts may di er in their ability to correct for miscategorization. This is what we investigate in columns (2)-(7). In column (2), we look at analyst experience. We measure experience as the number of years an analyst appears in the I/B/E/S database. The interaction term is positive and signi cant at 10%; This is weakly consistent with the view that more experienced analysts rely less on the \o cial" category when o cial and fundamental consensuses diverge. In column (3), we interact the absolute industry di erence with rm-level experience of the analyst de ned as the number of years analyst i has been issuing forecasts for rm j. Again, we nd a statistically signi cant positive sign. These rst two results suggest that experience reduces biases.

Next, we look at analyst's breadth of information. The e ect of following a broad set of stocks is ambiguous. On the one hand, it could create cognitive overload, making

it more difficult to correct categorization bias. On the other hand, it could provide the analyst with adequate benchmarks to form more rational expectations. In column (4), we interact official-fundamental divergence with the number of stocks covered by the analyst. A broader view seems to be beneficial: We find a statistically significantly positive sign for the interaction effect between the absolute industry difference and the number of stocks covered by analyst i in year t. Similarly, analysts who cover stocks from a higher number of official industries also tend to be more rational when a firm appears to be only limitedly representative of its official sector (see column (5)). Covering more official industries indicates to some extent that an analyst's understanding of comparable firms depends to a lesser extent on official industry classifications. In column (6), we seek to capture whether an analyst is more focused on the official or fundamental peers of firm j. We expect that, when an analyst follows many fundamental peers of a firm, she will form more rational expectations, in particular when it is important to do so. To implement this, we calculate the number of official peers $\#\ Official_{i,j,t}$, as the number of firms belonging to the same SIC2 industry as firm j that are also covered by analyst i in year t. Similarly, the number of fundamental peers $\#\ Fundamental_{i,j,t}$ is the number of firms covered by analyst i that are linked to firm j in the Hoberg and Phillips sense. We normalize the difference between official and fundamental peers by the number of stocks covered by the analyst. We label this measure $Coverage_{i,j,t}$ = $(\#\ Official_{i,j,t} - \#\ Fundamental_{i,j,t})/Stocks_{i,j,t}$. We expect that whenever an analyst's stock universe is tilted more toward official peers (positive values for $Coverage$), she will not deviate enough from official industry averages when in fact such deviation would be rational (high absolute industry difference). In line with this argument we find a negative coefficient which is highly statistically significant for the interaction between $Coverage$ and $|FP(T)_{j,official,t} - FP(T)_{j,fundamental,t}|$ (see column (6)).

Finally, we examine whether the size of the brokerage house at which the analyst is employed has an impact on excessive industry anchoring. We find no evidence of such an effect (see column (7)). Altogether, our results suggest that analysts with more informa-

tion (more experience, following a broader set stock, following more relevant comparables) tend to be less subject to categorization bias.

V Conclusion

This paper is part of the growing literature on the \economics of inattention" that studies the real e ects induced by cognitive costs of information processing. In the rst part of the paper, we explore the stock-price e ects of industry categorization bias. If some investors mentally group rms belonging to the same o cial industry category when processing new information, one should expect over-reaction to o cial industry shocks and under-reaction to fundamental shocks that are not captured by the o cial industry classi cation. To test this hypothesis, we compute the di erence between returns of a rm's o cial industry and a portfolio of its fundamentals peers, which we de ne according to whether rms operate in similar product markets (see Hoberg and Phillips (2010a,b)). In line with the existence of industry categorization bias, we nd that, in the short run rms comove strongly with their o cial industry and weakly with their fundamental peers, whereas this pattern reverses at longer horizons. Furthermore, by constructing portfolios based on this insight, we show that divergence between a rm's o cial industry returns and those of its fundamental peers strongly predicts a rm's subsequent reversal toward fundamentals. The long-short strategy based on this signal generates highly signi cant risk adjusted excess returns, which become even larger when conditioning on whether a rm's stock price has moved closely in line with its o cial industry in the previous month. We thus conclude that the overemphasis of o cial industry classi cations by boundedly rational investors creates large and predictable short-term deviations of prices from fundamentals, resulting in cross sectional return predictability.

Second, by studying analyst forecasts errors, we explore the extent to which nancial analysts are themselves subject to industry categorization biases. We provide evidence that analysts tend to overemphasize o cial industry categories, resulting in cross sectional

predictability of earnings forecast errors for firms which are poorly representative of their official industry.

References

Barberis, N., A. Shleifer, and J. Wurgler, 2005, Comovement, *Journal of Financial Economics* 75, 283–317.

Carhart, Marc M., 1997, On Persistence in Mutual Fund Performance, *Journal of Finance* 52, 57–82.

Cen, L., G. Hilary, and K.C. Wei, 2011, The role of anchoring bias in the equity market: Evidence from analysts' earnings forecasts and stock returns, forthcoming in *Journal of Financial and Quantitative Analysis*.

Chan, L.K.C., J. Lakonishok, and B. Swaminathan, 2007, Industry classifications and return comovement, *Financial Analysts Journal* pp. 56–70.

Chan, W.S., 2003, Stock price reaction to news and no-news: drift and reversal after headlines, *Journal of Financial Economics* 70, 223–260.

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

Cohen, Lauren, and Andrea Frazzini, 2008, Economic Links and Predictable Returns, *The Journal of Finance* 63, 1977–2011.

Cohen, Lauren, and Dong Lou, 2010, Complicated firms, forthcoming in the *Journal of Financial Economics*.

Da, Z., J. Engelberg, and P. Gao, 2011, In search of attention, *The Journal of Finance* 66, 1461–1499.

Daniel, K., M. Grinblatt, S. Titman, and R. Wermers, 1997, Measuring mutual fund performance with characteristic-based benchmarks, *Journal of Finance* pp. 1035{1058.

Diether, K.B., C.J. Malloy, and A. Scherbina, 2002, Di erences of opinion and the cross section of stock returns, *The Journal of Finance* 57, 2113{2141.

Fama, E.F., and J.D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *The Journal of Political Economy* pp. 607{636.

Fama, Eugene F., and Kenneth French, 1993, Common risk factors in the returns on bonds and stocks, *Journal of Financial Economics* 33, 3{53.

Foster, G., 1981, Intra-industry information transfers associated with earnings releases, *Journal of Accounting and Economics* 3, 201{232.

Gabaix, Xavier, 2010, A Sparsity-Based Model of Bounded Rationality, Working Paper, Stern School of Business, New York University.

Guenther, D.A., and A.J. Rosman, 1994, Di erences between compustat and crsp sic codes and related e ects on research, *Journal of Accounting and Economics* 18, 115{128.

Hoberg, G., and G. Phillips, 2010a, Product market synergies and competition in mergers and acquisitions: A text-based analysis, *Review of Financial Studies* 23, 3773.

Hoberg, Gerard, and Gordon Phillips, 2010b, Text-Based Network Industries and Endogenous Product Di erentiation, University of Maryland Working Paper.

Hong, H., and M. Kacperczyk, 2010, Competition and bias, *The Quarterly Journal of Economics* 125, 1683.

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

Hong, H., T. Lim, and J.C. Stein, 2000, Bad news travels slowly: Size, analyst coverage, and the pro tability of momentum strategies, *The Journal of Finance* 55, 265{295.

Hong, H., J. Stein, and J. Yu, 2007, Simple forecasts and paradigm shifts, *Journal of Finance*.

Hong, H., W. Torous, and R. Valkanov, 2007, Do industries lead stock markets?, *Journal of Financial Economics* 83, 367{396.

Hou, K., 2007, Industry information di usion and the lead-lag e ect in stock returns, *Review of Financial Studies* 20, 1113.

——— , and T.J. Moskowitz, 2005, Market frictions, price delay, and the cross-section of expected returns, *Review of Financial Studies* 18, 981{1020.

Ivkovic, Z., and N. Jegadeesh, 2004, The timing and value of forecast and recommendation revisions, *Journal of Financial Economics* 73, 433{463.

Jegadeesh, N., 1990, Evidence of predictable behavior of security returns, *Journal of Finance* 45, 881{98.

Kahle, K.M., and R.A. Walkling, 1996, The impact of industry classi cations on nancial research, *Journal of Financial and Quantitative Analysis* 31, 309{335.

Lo, A.W., and A.C. MacKinlay, 1990, When are contrarian pro ts due to stock market overreaction?, *Review of Financial studies* 3, 175.

Malloy, C.J., 2005, The geography of equity analysis, *The Journal of Finance* 60, 719{755.

Menzly, L., and O. Ozbas, 2006, Cross-industry momentum, manuscript, University of Southern California.

Moskowitz, T.J., and M. Grinblatt, 1999, Do industries explain momentum?, *The Journal of Finance* 54, 1249{1290.

Mullainathan, S., 2002, Thinking through categories, *mimeo MIT*.

——— , A. Shleifer, and J. Schwartzstein, 2008, Coarse thinking and persuasion, *Quarterly Journal of Economics* pp. 577{619.

Payne, J.L., and W.B. Thomas, 2003, The implications of using stock-split adjusted i/b/e/s data in empirical research, *Accounting Review* pp. 1049{1067.

Peng, L., and W. Xiong, 2006, Investor attention, overcon dence and category learning, *Journal of Financial Economics* 80, 563{602.

Petersen, M.A., 2009, Estimating standard errors in nance panel data sets: Comparing approaches, *Review of financial studies* 22, 435{480.

Ponti , J., 1996, Costly arbitrage: Evidence from closed-end funds, *The Quarterly Journal of Economics* 111, 1135.

Ramnath, S., 2002, Investor and analyst reactions to earnings announcements of related rms: An empirical analysis, *Journal of Accounting Research* 40, 1351{1376.

Robinson, D., and D. Glushkov, 2006, A note on ibes unadjusted data, Wharton Research Data Services.

Figures

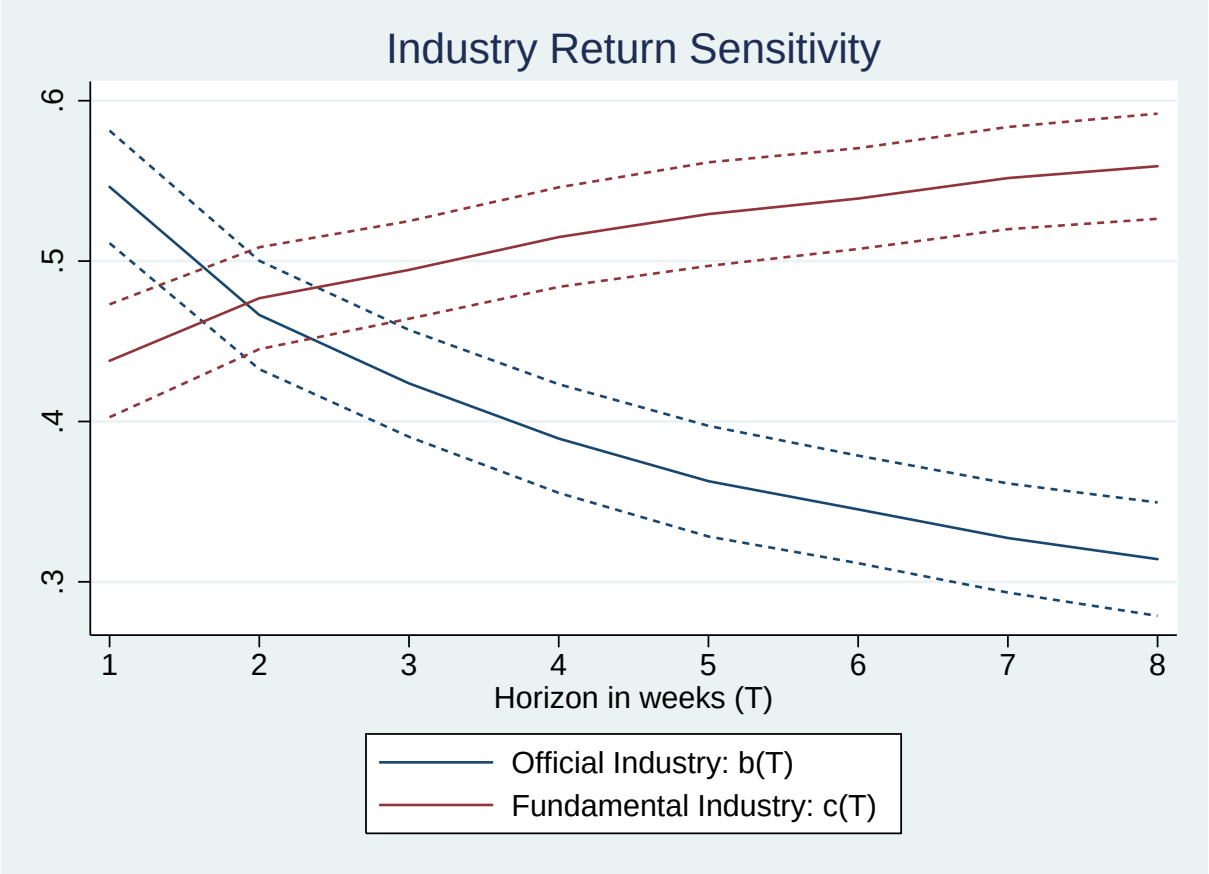

Figure 1. This figure shows how cumulative stock returns covary with official and fundamental cumulative industry returns. A firm's fundamental industry return is defined as the equally weighted return of a portfolio consisting of all firms that are linked to firm j in the Hoberg and Phillips (2010a,b) sense in the previous calendar year. A firm's official industry is defined as the historical two digit SIC code reported by Compustat for the previous calendar year. We regress a firm's T week cumulative return on the firm's fundamental and the firm's official T week cumulative industry returns in a pooled cross section with double clustered standard errors (time and stock dimension). Formally, we estimate the following equation $r^T_{j;t} = a_T + b_T \times r^T_{j;official;t} + c_T \times r^T_{j;fundamental;t} + e^T_{j;t}$ for varying T. $r^T_{j;t}$ denotes firm $j's$ T week cumulative return. $r^T_{j;fundamental;t}$ and $r^T_{j;official;t}$ are the T week cumulative returns of baskets consisting of firms that are fundamentally and officially linked to firm j. We restrict the regressions to stocks for which the price at the beginning of the period exceeds \$5. In addition, we require the official and the fundamental industries to be populated by at least five different firms. We then plot the resulting coefficient estimates b_T and c_T against the time horizon T of the cumulative returns. The dotted lines show 95% confidence intervals for both coefficient estimates.

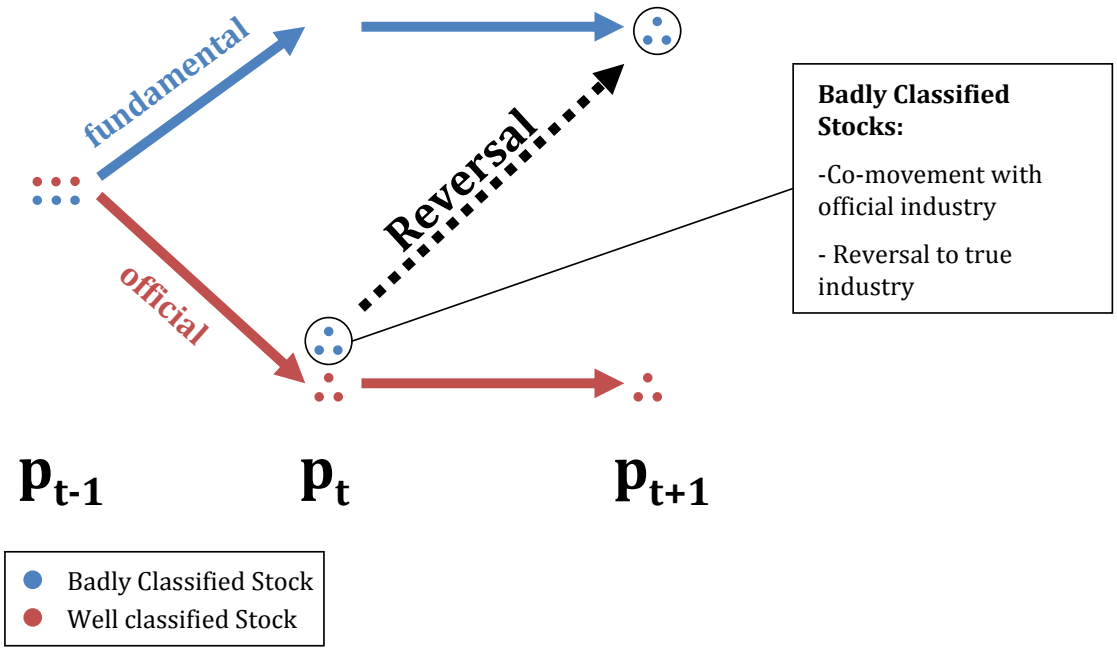

Figure 2. This diagram illustrates the idea behind the refined categorization bias strategy, which is based on a signal combining the official-fundamental industry return differential and a measure indicating whether a stock has moved closely in line with its official industry over the previous month.

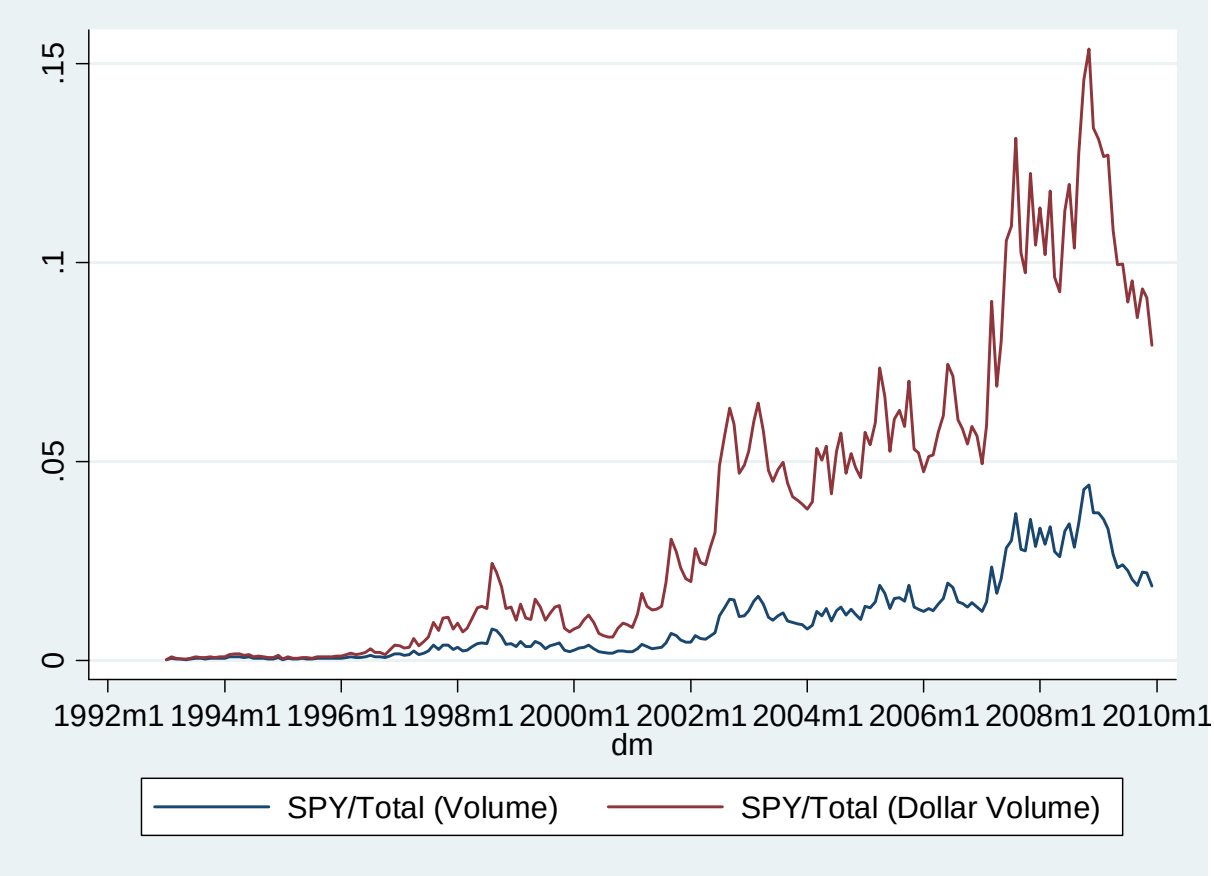

Figure 3. This figure plots monthly (dollar) volume of the exchange traded fund SPDR S&P 500 (Ticker: SPY) normalized by total (dollar) volume of the CRSP universe.

Tables

Table I

Summary Statistics: Returns

This table reports summary statistics for the returns sample, which runs from 1997-2009. All variables are defined in the Appendix. All variables that rely on Compustat data are trimmed at the first and 99th percentile.

	Mean	Median	SD	P25	P75	N
Panel A: Annual Variables						
SIC2 firms	224.786	140.000	188.681	66.000	372.000	48641
Hoberg and Phillips firms	83.278	34.000	118.682	11.000	98.000	47935
ln(BE/ME)	-0.723	-0.644	0.854	-1.200	-0.166	47218
Analysts	6.416	4.000	7.879	0.000	10.000	48641
NYSE Breakpoint	3.442	2.000	2.879	1.000	5.000	48641
Panel B: Monthly Variables						
$r_{j;t}$	0.012	0.000	0.201	-0.078	0.079	565266
$r_{j;official;t}$	0.011	0.012	0.084	-0.035	0.053	565266
$r_{j;fundamental;t}$	0.012	0.010	0.104	-0.039	0.057	564540
$\rho_{j;t}$	0.810	0.626	0.699	0.288	1.135	565261
$r_{j;official;t} - r_{j;fundamental;t}$	-0.001	0.001	0.068	-0.024	0.025	564540
$r_{j;t-2;t-12}$	0.089	0.004	0.597	-0.272	0.302	561824
$\ln(\mathrm{M})_{j;t-1}$	12.535	12.472	2.156	10.942	13.989	565266
Panel C: Weekly Variables						
$r_{j;t}$	0.003	-0.000	0.103	-0.038	0.036	2606105
$r_{j;official;t}$	0.003	0.004	0.036	-0.014	0.021	2606105
$r_{j;fundamental;t}$	0.003	0.003	0.043	-0.017	0.022	2606105
$\rho_{j;t}$	0.758	0.564	0.711	0.256	1.037	2379113
$r_{j;official;t} - r_{j;fundamental;t}$	0.000	0.000	0.024	-0.009	0.010	2606105

Table II

Summary Statistics: Analyst Forecasts

This table reports summary statistics of all employed firm, analyst and analyst-firm level variables. The sample period runs from 1997-2009. All variables are defined in the Appendix. All variables are trimmed at the first and 99th percentile.

	Mean	Median	SD	P25	P75	N
Panel A: Analyst-Firm Level						
$ForecastError(3)_{i;j;t}$	0.002	0.000	0.021	-0.006	0.009	114874
$FP(3)_{i;j;t}$	0.052	0.054	0.037	0.036	0.072	114874
$\rho(3)_{i;j;t}$	8.129	4.611	9.518	1.820	10.478	114874
# $\text{Official}_{i;j;t}$	4.433	3.000	4.737	1.000	6.000	108129
# $\text{Fundamental}_{i;j;t}$	3.569	2.000	3.935	1.000	5.000	108129
$\text{Coverage}_{i;j;t}$	0.058	0.000	0.191	0.000	0.143	108129
Firm $\text{Experience}_{i;j;t}$	3.572	2.000	3.756	1.000	5.000	114874
Panel B: Firm Level						
$ForecastError(0)_{j;t}$	-0.000	-0.000	0.005	-0.002	0.001	18751
$ForecastError(1)_{j;t}$	0.001	-0.000	0.009	-0.003	0.004	20487
$ForecastError(2)_{j;t}$	0.002	0.000	0.016	-0.004	0.008	20200
$ForecastError(3)_{j;t}$	0.004	0.001	0.022	-0.005	0.011	18550
$FP(0)_{j;t}$	0.050	0.053	0.044	0.033	0.071	18751
$FP(1)_{j;t}$	0.049	0.053	0.042	0.034	0.071	20487
$FP(2)_{j;t}$	0.053	0.056	0.043	0.036	0.074	20200
$FP(3)_{j;t}$	0.054	0.057	0.041	0.038	0.074	18550
$FP(0)_{j;official;t}$	0.048	0.048	0.025	0.030	0.064	18751
$FP(1)_{j;official;t}$	0.047	0.048	0.025	0.029	0.065	20487
$FP(2)_{j;official;t}$	0.050	0.051	0.026	0.032	0.066	20200
$FP(3)_{j;official;t}$	0.051	0.053	0.025	0.033	0.068	18550
$FP(0)_{j;fundamental;t}$	0.046	0.051	0.032	0.027	0.066	18751
$FP(1)_{j;fundamental;t}$	0.046	0.051	0.031	0.028	0.067	20487
$FP(2)_{j;fundamental;t}$	0.049	0.054	0.031	0.032	0.069	20200
$FP(3)_{j;fundamental;t}$	0.050	0.056	0.030	0.033	0.070	18550
$FP(0)_{j;official;t} - FP(0)_{j;fundamental;t}$	0.002	0.001	0.022	-0.006	0.012	18751
$FP(1)_{j;official;t} - FP(1)_{j;fundamental;t}$	0.001	0.000	0.021	-0.007	0.010	20487
$FP(2)_{j;official;t} - FP(2)_{j;fundamental;t}$	0.001	0.000	0.022	-0.007	0.010	20200
$FP(3)_{j;official;t} - FP(3)_{j;fundamental;t}$	0.001	-0.000	0.021	-0.007	0.009	18550
$\lvert FP(3)_{j;official;t} - FP(3)_{j;fundamental;t} \rvert$	0.014	0.008	0.015	0.003	0.020	18550
$\rho(0)_{j;t}$	0.549	0.364	0.608	0.157	0.716	18751
$\rho(1)_{j;t}$	0.538	0.360	0.601	0.154	0.703	20487
$\rho(2)_{j;t}$	0.540	0.359	0.596	0.153	0.706	20200
$\rho(3)_{j;t}$	0.542	0.364	0.601	0.156	0.707	18550
ln(Size)	13.728	13.570	1.714	12.509	14.778	22839
VolRet	0.122	0.105	0.073	0.071	0.153	22839
Ret	0.015	0.013	0.041	-0.007	0.034	22839
ln(BM)	-0.764	-0.706	0.693	-1.185	-0.299	22839
VolRoe	0.072	0.046	0.075	0.024	0.089	22839
Roe	0.181	0.186	0.209	0.100	0.277	22839
SP500	0.201	0.000	0.401	0.000	0.000	22839
ln(P)	3.098	3.135	0.735	2.588	3.606	22839
Analysts	6.811	5.000	6.156	2.000	9.000	22839
Panel C: Analyst Level						
$\text{Experience}_{i;t}$	5.986	4.000	5.317	2.000	9.000	27657
$\text{Industries}_{i;t}$	3.017	3.000	2.140	1.000	4.000	27657
$\text{Stocks}_{i;t}$	10.820	10.000	6.259	7.000	14.000	27657
Broker $\text{Size}_{i;t}$	53.297	38.000	46.796	16.000	81.000	27657

Table III

Weekly Categorization Bias Strategy (EW)

This table shows alphas and factor loadings from weekly performance regressions of the Categorization Bias Strategy. At the beginning of each week, stocks are sorted into five quintiles according to the difference between the stock's official and fundamental industry return during week $t-n$, i.e. $(r_{j;official;t-n} - r_{j;fundamental;t-n})$. A firm's official industry return is defined as the equally weighted return of all its SIC2 peers. A firm's fundamental industry return is defined as the equally weighted return of a portfolio consisting of all its Hoberg and Phillips peers. We consider a maximum horizon of six weeks ($n = 6$). Portfolios in the Q1 portfolio have experienced the most negative difference between official and fundamental industry returns. Q1-Q5 is a portfolio that is long in stocks for which the return difference between the official and the fundamental industry portfolios has been the most negative (Q1, bottom 20%) and short in stocks for which the return differential has been the most positive (Q5, top 20%). All portfolios are equally weighted and are rebalanced on a weekly basis.

	alpha	mktrf	smb	hml	mom	R2	N
Panel A: Signal (t-1)							
Q1	0.281***	0.935***	0.790***	0.166***	-0.112***	0.903	675
	(7.18)	(61.84)	(29.37)	(6.39)	(-7.13)		
Q5	-0.0918**	1.037***	0.717***	0.0663**	-0.0852***	0.914	675
	(-2.33)	(68.16)	(26.47)	(2.54)	(-5.41)		
Q1-Q5	0.373***	-0.101***	0.0739	0.0995**	-0.0265	0.0401	675
	(5.66)	(-3.98)	(1.63)	(2.28)	(-1.01)		
Panel B: Signal (t-2)							
Q1	0.173***	0.948***	0.741***	0.202***	-0.0981***	0.898	674
	(4.30)	(61.01)	(26.82)	(7.57)	(-6.10)		
Q5	0.00627	1.047***	0.741***	0.0591**	-0.107***	0.911	674
	(0.15)	(66.38)	(26.41)	(2.19)	(-6.55)		
Q1-Q5	0.167**	-0.0989***	0.000349	0.143***	0.00885	0.0493	674
	(2.45)	(-3.76)	(0.01)	(3.16)	(0.32)		
Panel C: Signal (t-3)							
Q1	0.210***	0.978***	0.789***	0.176***	-0.0704***	0.907	673
	(5.32)	(64.33)	(29.16)	(6.77)	(-4.47)		
Q5	-0.000617	1.005***	0.689***	0.0335	-0.156***	0.905	673
	(-0.01)	(63.04)	(24.29)	(1.23)	(-9.45)		
Q1-Q5	0.210***	-0.0278	0.0993**	0.143***	0.0858***	0.0347	673
	(3.07)	(-1.05)	(2.11)	(3.15)	(3.13)		
Panel D: Signal (t-4)							
Q1	0.143***	0.984***	0.780***	0.199***	-0.0768***	0.904	672
	(3.55)	(63.35)	(28.21)	(7.47)	(-4.77)		
Q5	0.0457	1.004***	0.735***	0.0575**	-0.157***	0.906	672
	(1.10)	(62.75)	(25.80)	(2.10)	(-9.46)		
Q1-Q5	0.0971	-0.0202	0.0451	0.141***	0.0800***	0.0271	672
	(1.40)	(-0.76)	(0.95)	(3.09)	(2.90)		
Panel E: Signal (t-5)							
Q1	0.112***	1.012***	0.899***	0.147***	-0.0281*	0.927	671
	(3.10)	(72.27)	(36.08)	(6.14)	(-1.94)		
Q5	0.0600	0.971***	0.678***	0.0944***	-0.204***	0.909	671
	(1.52)	(63.55)	(24.92)	(3.61)	(-12.90)		
Q1-Q5	0.0522	0.0401*	0.221***	0.0526	0.176***	0.119	671
	(0.84)	(1.67)	(5.19)	(1.29)	(7.11)		
Panel F: Signal (t-6)							
Q1	0.165***	1.029***	0.834***	0.101***	-0.00889	0.918	670
	(4.28)	(69.03)	(31.42)	(3.98)	(-0.58)		
Q5	0.0293	0.960***	0.669***	0.112***	-0.208***	0.897	670
	(0.70)	(59.20)	(23.16)	(4.02)	(-12.41)		
Q1-Q5	0.136**	0.0687***	0.165***	-0.0102	0.199***	0.119	670
	(2.03)	(2.66)	(3.58)	(-0.23)	(7.45)		

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table IV

Monthly Categorization Bias Strategy (VW): By Market Capitalization

At the beginning of each month, stocks are ranked according to their market capitalization and separated into brackets of small, mid and large capitalization stocks (three terciles). Within each tercile, stocks are then sorted into five quintiles according to the difference between an equally weighted portfolio of stocks belonging to the firm's official and an equally weighted portfolio consisting of stocks belonging to a firm's fundamental industry ($r_{j;official;t-1} - r_{j;fundamental;t-1}$). Within each size tercile, the ranked stocks are assigned to one of five quintile portfolios. Q1 contains firms for which the return differential between the official and the fundamental industry is the lowest. All stocks are value weighted within a quintile portfolio and rebalanced every month. Monthly alphas as well as Fama and French (1993) and Carhart (1997) factor loadings are in percent. Panels A, B and C show the results for value weighted portfolios consisting of small, mid and large stocks respectively.

	alpha	mktrf	smb	hml	mom	R2	N
Panel A: Small Cap Stocks (First Tercile)							
Q1	0.915***	0.808***	1.157***	0.211***	-0.116***	0.835	155
	(3.69)	(14.41)	(17.32)	(2.91)	(-2.73)		
Q2	0.0683	0.839***	0.894***	0.229***	-0.0994***	0.871	155
	(0.36)	(19.28)	(17.23)	(4.06)	(-3.02)		
Q3	-0.0535	0.827***	0.815***	0.416***	-0.135***	0.817	155
	(-0.24)	(16.32)	(13.50)	(6.34)	(-3.53)		
Q4	-0.420*	0.979***	0.820***	0.416***	-0.112***	0.852	155
	(-1.92)	(19.82)	(13.93)	(6.50)	(-2.99)		
Q5	-0.629***	1.025***	0.903***	0.328***	-0.144***	0.844	155
	(-2.58)	(18.57)	(13.73)	(4.59)	(-3.46)		
Q1-Q5	1.544***	-0.217***	0.254***	-0.117	0.0284	0.103	155
	(4.25)	(-2.64)	(2.60)	(-1.10)	(0.46)		
Panel B: Mid Cap Stocks (Second Tercile)							
Q1	0.652**	0.955***	1.133***	0.0854	-0.0520	0.807	155
	(2.21)	(14.30)	(14.23)	(0.99)	(-1.03)		
Q2	0.0130	0.937***	0.720***	0.326***	-0.0270	0.887	155
	(0.08)	(24.33)	(15.67)	(6.52)	(-0.93)		
Q3	-0.241	0.941***	0.603***	0.508***	-0.0293	0.822	155
	(-1.15)	(19.80)	(10.65)	(8.25)	(-0.82)		
Q4	-0.350*	1.014***	0.702***	0.312***	-0.0692**	0.878	155
	(-1.84)	(23.56)	(13.69)	(5.59)	(-2.13)		
Q5	-0.536**	1.202***	0.749***	0.225***	-0.0760*	0.831	155
	(-1.98)	(19.66)	(10.28)	(2.84)	(-1.65)		
Q1-Q5	1.187**	-0.246**	0.384***	-0.139	0.0240	0.0946	155
	(2.39)	(-2.19)	(2.87)	(-0.95)	(0.28)		
Panel C: Large Cap Stocks (Third Tercile)							
Q1	0.255	0.903***	0.0334	-0.153**	-0.0190	0.768	155
	(1.17)	(18.28)	(0.57)	(-2.39)	(-0.51)		
Q2	0.144	0.872***	-0.164***	0.0519	0.0305	0.858	155
	(1.03)	(27.55)	(-4.34)	(1.26)	(1.28)		
Q3	-0.103	0.961***	-0.0923*	0.277***	0.0296	0.818	155
	(-0.58)	(24.03)	(-1.94)	(5.35)	(0.98)		
Q4	-0.216	0.968***	-0.179***	0.0236	0.0220	0.870	155
	(-1.45)	(28.58)	(-4.44)	(0.54)	(0.86)		
Q5	0.00746	1.060***	-0.124**	-0.164**	-0.0573	0.814	155
	(0.03)	(21.41)	(-2.10)	(-2.56)	(-1.53)		
Q1-Q5	0.247	-0.157*	0.158	0.0112	0.0383	0.0467	155
	(0.66)	(-1.86)	(1.57)	(0.10)	(0.60)		

t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table V

Monthly Categorization Bias Strategy (EW): By Industry Follower

At the beginning of each month, stocks are ranked according to whether their returns tracked the returns of their official industries over the previous month. To do so, we calculate the absolute value of the difference between the firm's return and the firm's official industry return normalized by the rolling standard deviation of the difference. We denote this measure as $\rho_{j;t-1} = \left|r_{j;t-1} - r_{j;official;t-1}\right| / \sigma_{j;24}(r_{j;t-1} - r_{j;official;t-1})$. Low values of $rho_{j;t-1}$ indicate that over the prior month, the return of firm j has been abnormally close to the return of the official industry, while high values indicate divergence. Stocks in the first tercile of the distribution of ρ are labled as *Strong O cial Industry Follower* stocks. In each *O cial Industry Follower* tercile, we then sort stocks according to the difference between an equally weighted portfolio of stocks belonging to the firm's official and an equally weighted portfolio consisting of stocks that belong to the firm's fundamental industry ($r_{j;official;t-1} - r_{j;fundamental;t-1}$). Q1 contains firms for which the return differential between the official and the fundamental industry is the lowest. Q1-Q5 is a portfolio that is long in stocks for which the return difference between the official and the fundamental industry portfolios has been the most negative (Q1, bottom 20%) and short in stocks for which the return differential has been the most positive (Q5, top 20%). All stocks are equally weighted within a quintile portfolio and rebalanced every month. Monthly alphas as well as Fama and French (1993) and Carhart (1997) factor loadings are in percent. Panels A, B and C show the results for quintile portfolios restricted to *Strong, Medium and Weak O cial Industry Follower* stocks (three terciles of ρ).

	alpha	mktrf	smb	hml	mom	R2	N
Panel A: Strong O cial Industry Follower (First Tercile)							
Q1	0.957***	0.891***	0.902***	0.0857	-0.0776**	0.842	155
	(4.18)	(17.20)	(14.61)	(1.28)	(-1.99)		
Q2	0.0738	0.904***	0.564***	0.253***	-0.0464*	0.890	155
	(0.48)	(25.72)	(13.46)	(5.55)	(-1.75)		
Q3	-0.401**	0.913***	0.495***	0.381***	-0.0227	0.878	155
	(-2.54)	(25.60)	(11.65)	(8.24)	(-0.84)		
Q4	-0.424**	0.987***	0.574***	0.249***	-0.0364	0.870	155
	(-2.32)	(23.80)	(11.61)	(4.63)	(-1.16)		
Q5	-0.617***	1.060***	0.649***	0.0382	-0.0967**	0.860	155
	(-2.81)	(21.28)	(10.94)	(0.59)	(-2.57)		
Q1-Q5	1.575***	-0.169*	0.253**	0.0475	0.0191	0.0615	155
	(4.10)	(-1.95)	(2.44)	(0.42)	(0.29)		
Panel B: Medium O cial Industry Follower (Second Tercile)							
Q1	0.540**	0.921***	0.835***	0.112*	-0.0752**	0.860	155
	(2.57)	(19.37)	(14.74)	(1.82)	(-2.09)		
Q2	0.320**	0.878***	0.545***	0.221***	-0.0459*	0.892	155
	(2.13)	(25.82)	(13.45)	(5.02)	(-1.79)		
Q3	0.0326	0.872***	0.468***	0.392***	-0.0392	0.839	155
	(0.18)	(21.69)	(9.77)	(7.52)	(-1.29)		
Q4	-0.452***	0.972***	0.548***	0.307***	-0.0437	0.879	155
	(-2.64)	(25.10)	(11.86)	(6.11)	(-1.50)		
Q5	-0.423*	1.096***	0.642***	0.180***	-0.105***	0.845	155
	(-1.81)	(20.69)	(10.17)	(2.62)	(-2.64)		
Q1-Q5	0.963**	-0.175**	0.193*	-0.0676	0.0303	0.0586	155
	(2.51)	(-2.02)	(1.86)	(-0.60)	(0.46)		
Panel C: Weak O cial Industry Follower (Third Tercile)							
Q1	0.469	0.820***	1.041***	0.0645	-0.0218	0.760	155
	(1.56)	(12.04)	(12.83)	(0.73)	(-0.42)		
Q2	0.0326	0.889***	0.602***	0.268***	-0.0475*	0.893	155
	(0.21)	(25.61)	(14.55)	(5.97)	(-1.81)		
Q3	-0.217	0.873***	0.494***	0.521***	-0.0679**	0.857	155
	(-1.28)	(22.67)	(10.77)	(10.46)	(-2.34)		
Q4	-0.162	0.973***	0.516***	0.412***	-0.0919***	0.903	155
	(-1.07)	(28.31)	(12.62)	(9.25)	(-3.54)		
Q5	-0.300	1.070***	0.562***	0.323***	-0.130***	0.857	155
	(-1.41)	(22.27)	(9.82)	(5.19)	(-3.58)		
Q1-Q5	0.769*	-0.250**	0.479***	-0.259**	0.108	0.204	155
	(1.73)	(-2.49)	(3.99)	(-1.98)	(1.42)		

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table VI

Monthly Categorization Bias Strategy (VW): Strong Industry Follower; By Market Capitalization

In this table, we restrict the portfolios to *Strong Official Industry Follower Stocks* only (low values of ρ). See the previous table for a formal definition of ρ. In each month, we first sort the set of *Strong Official Industry Follower Stocks* into three size brackets (small, mid and large caps). Within each size tercile, we then sort the stocks according to the difference between an equally weighted portfolio of stocks belonging to the stock's official and an equally weighted portfolio consisting of stocks that belong to the stock's fundamental industry ($r_{j;official;t-1} - r_{j;fundamental;t-1}$). The ranked stocks are assigned to one of five quintile portfolios. Q1 contains firms for which the return differential between the official and the fundamental industry is the lowest. Q1-Q5 is a portfolio that is long in stocks for which the return difference between the official and the fundamental industry portfolios has been the most negative (Q1, bottom 20%) and short in stocks for which the return differential has been the most positive (Q5, top 20%). All stocks are value weighted within a quintile portfolio and rebalanced every month. Monthly alphas as well as Fama and French (1993) and Carhart (1997) factor loadings are in percent. Panels A, B and C show alphas and loadings for value weighted portfolios consisting of small, mid and large stocks respectively.

	alpha	mktrf	smb	hml	mom	R2	N
Panel A: Small Cap Stocks (First Tercile)							
Q1	1.338***	0.834***	1.214***	0.240**	-0.103*	0.762	155
	(4.15)	(11.43)	(13.96)	(2.54)	(-1.88)		
Q2	-0.0419	0.856***	0.778***	0.175**	-0.0764	0.741	155
	(-0.15)	(13.37)	(10.20)	(2.11)	(-1.58)		
Q3	-0.407	0.837***	0.803***	0.337***	-0.139***	0.749	155
	(-1.47)	(13.35)	(10.75)	(4.15)	(-2.93)		
Q4	-0.544**	0.993***	0.816***	0.322***	-0.0724	0.785	155
	(-1.98)	(15.96)	(11.00)	(4.00)	(-1.54)		
Q5	-0.731**	0.948***	0.983***	0.196**	-0.192***	0.772	155
	(-2.33)	(13.36)	(11.63)	(2.14)	(-3.59)		
Q1-Q5	2.069***	-0.113	0.231*	0.0434	0.0888	0.0504	155
	(4.72)	(-1.14)	(1.95)	(0.34)	(1.18)		
Panel B: Mid Cap Stocks (Second Tercile)							
Q1	0.760**	0.947***	1.101***	0.110	-0.0637	0.774	155
	(2.37)	(13.07)	(12.75)	(1.17)	(-1.16)		
Q2	-0.235	0.936***	0.771***	0.385***	-0.0346	0.849	155
	(-1.15)	(20.20)	(13.97)	(6.42)	(-0.99)		
Q3	-0.189	0.968***	0.621***	0.398***	-0.0438	0.791	155
	(-0.78)	(17.73)	(9.55)	(5.63)	(-1.06)		
Q4	-0.393	0.984***	0.726***	0.258***	-0.0294	0.794	155
	(-1.54)	(17.03)	(10.55)	(3.45)	(-0.67)		
Q5	-0.849***	1.209***	0.749***	0.0616	-0.0577	0.805	155
	(-2.82)	(17.72)	(9.21)	(0.70)	(-1.12)		
Q1-Q5	1.609***	-0.262**	0.352**	0.0480	-0.00605	0.0638	155
	(3.05)	(-2.20)	(2.48)	(0.31)	(-0.07)		
Panel C: Large Cap Stocks (Third Tercile)							
Q1	0.724**	0.961***	0.191**	-0.0403	0.00996	0.653	155
	(2.36)	(13.83)	(2.31)	(-0.45)	(0.19)		
Q2	0.206	0.955***	-0.185***	0.124*	0.0897**	0.679	155
	(0.82)	(16.85)	(-2.74)	(1.69)	(2.10)		
Q3	-0.606***	0.935***	-0.148**	0.275***	0.0224	0.737	155
	(-2.78)	(18.93)	(-2.51)	(4.30)	(0.60)		
Q4	-0.0231	1.018***	-0.0758	-0.0679	0.0447	0.786	155
	(-0.11)	(20.74)	(-1.30)	(-1.07)	(1.20)		
Q5	0.0884	1.036***	-0.0454	-0.225***	0.00817	0.707	155
	(0.31)	(15.86)	(-0.58)	(-2.65)	(0.17)		
Q1-Q5	0.636	-0.0759	0.237**	0.184	0.00179	0.0340	155
	(1.43)	(-0.76)	(1.98)	(1.42)	(0.02)		

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table VII

Pooled Cross-Sectional Regressions (Monthly Returns)

This table shows results from pooled cross sectional forecast regressions with double clustered standard errors (month-firm). The dependent variable is the monthly stock return of firm j. We restrict the regressions to stock-month observations for which the stock price at the end of the preceding month is greater or equal to \$5. All specifications include month dummies. *Weak, Med* and *Strong Ind Follower*$_{j;t-1}$ indicate terciles of the *O cial Industry Follower* measure ρ. Equations (3), (4), (5) and (6) are restricted to *Strong O cial Industry Follower* stocks. *Small, Mid and Large Cap*$_{j;t-1}$ are dummy variables indicating size terciles based on market capitalization in each month. *Med and High Analyst*$_{j;t-1}$ indicate whether the stock receives medium or high analyst coverage. *After 2002* is a dummy equal to one for firm-month observations after 2002. All other variables are defined in the appendix.

	(1)	(2)	(3)	(4)	(5)	(6)
$r_{j;official;t-1} - r_{j;fundamental;t}$						

Table VIII

Monthly Pseudo Industry Classification Strategy (EW)

In this table, we construct a portfolio strategy based on a Pseudo Hoberg and Phillips Industry classification (see section III.3 for details). We first sort firms into terciles of the previously defined Industry Follower Measure $\rho_{j;t-1}$ (see the appendix for a formal definition of $\rho_{j;t-1}$). Then, within each tercile of $\rho - j, t-1$, we sort firms according to $r_{j;official;t-1} - r_{j;pseudo;t-1}$. $r_{j;pseudo;t-1}$ is the return on an equally weighted portfolio consisting of random firms. For firm j, the number of random firms contained in the pseudo portfolio is equal to the number of firm $j's$ Hoberg and Phillips links in that month. As in preceding portfolio sorts, $r_{j;official;t-1}$ is an equally weighted portfolio made up of all firms belonging to firm $j's$ two digit SIC code. Q1 contains portfolios for which the firm's official-pseudo industry return differential is most negative. Panels A, B and C show the results for quintile portfolios restricted to *Strong, Medium and Weak O cial Industry Follower* stocks (three terciles of ρ).

	alpha	mktrf	smb	hml	mom	R2	N
Panel A: Strong O cial Industry Follower (First Tercile)							
Q1	-0.687***	1.041***	0.511***	0.253***	-0.127***	0.793	155
	(-2.68)	(17.92)	(7.38)	(3.37)	(-2.89)		
Q2	-0.397*	1.053***	0.523***	0.214***	-0.0603*	0.855	155
	(-1.93)	(22.67)	(9.45)	(3.56)	(-1.72)		
Q3	-0.0149	0.920***	0.682***	0.129***	-0.0510**	0.915	155
	(-0.10)	(27.62)	(17.18)	(2.99)	(-2.03)		
Q4	0.201	0.873***	0.707***	0.172***	-0.0630**	0.881	155
	(1.17)	(22.42)	(15.24)	(3.40)	(-2.15)		
Q5	0.483	0.852***	0.889***	0.118	0.0288	0.738	155
	(1.64)	(12.73)	(11.15)	(1.36)	(0.57)		
Q1-Q5	-1.170**	0.189*	-0.378***	0.135	-0.155*	0.125	155
	(-2.35)	(1.68)	(-2.81)	(0.93)	(-1.82)		
Panel B: Medium O cial Industry Follower (Second Tercile)							
Q1	-0.511**	1.046***	0.484***	0.356***	-0.130***	0.837	155
	(-2.33)	(21.07)	(8.18)	(5.53)	(-3.47)		
Q2	-0.0863	1.023***	0.568***	0.188***	-0.0799**	0.851	155
	(-0.41)	(21.66)	(10.10)	(3.07)	(-2.24)		
Q3	0.0743	0.931***	0.621***	0.230***	-0.0378	0.912	155
	(0.52)	(28.50)	(15.95)	(5.42)	(-1.54)		
Q4	0.264*	0.864***	0.647***	0.189***	-0.0597**	0.895	155
	(1.72)	(24.77)	(15.58)	(4.18)	(-2.27)		
Q5	0.318	0.854***	0.790***	0.222***	0.000292	0.752	155
	(1.19)	(14.10)	(10.95)	(2.82)	(0.01)		
Q1-Q5	-0.829*	0.192**	-0.306***	0.134	-0.130*	0.127	155
	(-1.93)	(1.97)	(-2.64)	(1.06)	(-1.78)		
Panel C: Weak O cial Industry Follower (Third Tercile)							
Q1	-0.622***	1.022***	0.528***	0.459***	-0.112***	0.874	155
	(-3.36)	(24.41)	(10.59)	(8.46)	(-3.56)		
Q2	-0.201	0.964***	0.557***	0.268***	-0.0999***	0.870	155
	(-1.10)	(23.31)	(11.29)	(5.00)	(-3.20)		
Q3	0.150	0.925***	0.597***	0.309***	-0.0843***	0.916	155
	(1.08)	(29.32)	(15.89)	(7.55)	(-3.54)		
Q4	0.248	0.829***	0.636***	0.306***	-0.0833***	0.868	155
	(1.47)	(21.76)	(13.99)	(6.20)	(-2.90)		
Q5	0.238	0.911***	0.724***	0.374***	0.00643	0.790	155
	(1.00)	(16.93)	(11.29)	(5.37)	(0.16)		
Q1-Q5	-0.860**	0.110	-0.196**	0.0845	-0.119*	0.0913	155
	(-2.35)	(1.33)	(-1.99)	(0.79)	(-1.90)		

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table IX

Firm-Level Forecast Error

The dependent variable in columns (1)–(4) is the average firm-level forecast error associated with forecasts issued in the first fiscal quarter of firm j (i.e. forecasts with a horizon of three quarters). In calculating the average firm level forecast error, we use the last forecast in quarter t that analyst i issues or revises for firm $j's$ year end earnings. The dependent variables in columns (5), (6) and (7) are the average firm-level forecast errors associated with the last forecast issued or revised in the second, third and fourth quarter of the fiscal year respectively. Thus, the forecast horizons in these regressions are two, one or zero quarters. $FP(T)_{j;official;t}$ is the average forecast to price ratio prevailing in firm $j's$ two digit SIC industry in quarter t. $FP(T)_{j;fundamental;t}$ is the average forecast to price ratio prevailing among firm $j's$ Hoberg and Phillips peers. $\rho(T)_{j;t}$ measures whether firm $j's$ average forecast to price ratio in quarter t is close to that of its official industry and is formally defined as $\rho(T)_{j;t} = |FP(T)_{j;t} - FP(T)_{j;official;t}|/\sigma_t(FP(T)_{j;t} - FP(T)_{j;official;t})$ where $FP(T)_{j;t}$ is the average forecast to price ratio for firm j in year t and σ_t denotes the cross-sectional standard deviation in year t. All other variables are defined in the appendix. All regressions include year fixed effects. We restrict all regressions to forecasrs for which for which the stock price of the firm at the beginning of the quarter exceeds $5 and for which the industry portfolios are populated by at least five different firms. Standard errors are clustered at the firm level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$FP(T)_{j;official;t} - FP(T)_{j;fundamental;t}$	0.0034	-0.0137	0.0437***	0.0292**	0.0013	-0.0070	-0.0011
	(0.43)	(-1.62)	(3.70)	(2.26)	(0.15)	(-1.39)	(-0.42)
$\rho(T)_{j;t}$			0.0026***	0.0008*	0.0004	0.0003**	0.0004***
			(7.28)	(1.79)	(1.37)	(1.98)	(4.09)
$\rho(T)_{j;t} \times (FP(T)_{j;official;t} - FP(T)_{j;fundamental;t})$			-0.0489***	-0.0493***	-0.0226**	-0.0042	-0.0013
			(-4.13)	(-3.48)	(-2.26)	(-0.72)	(-0.47)
ln(Size)		-0.0000		0.0000	0.0004***	0.0004***	0.0002***
		(-0.02)		(0.03)	(2.79)	(4.27)	(3.55)
VolRet		0.0246***		0.0232***	0.0111***	0.0009	-0.0021**
		(6.46)		(6.08)	(3.98)	(0.55)	(-2.49)
Ret		-0.0573***		-0.0567***	-0.0235***	-0.0088***	-0.0030**
		(-9.42)		(-9.24)	(-5.54)	(-3.56)	(-2.35)
ln(BM)		0.0003		0.0001	0.0001	0.0002	-0.0003***
		(0.73)		(0.23)	(0.40)	(1.46)	(-3.50)
VolRoe		-0.0016		-0.0024	-0.0000	-0.0021*	-0.0030***
		(-0.52)		(-0.76)	(-0.00)	(-1.79)	(-4.63)
Roe		-0.0012		-0.0019	-0.0008	0.0001	0.0003
		(-0.96)		(-1.45)	(-0.90)	(0.16)	(1.07)
SP500		0.0003		0.0003	0.0001	-0.0002	-0.0002*
		(0.61)		(0.62)	(0.28)	(-0.81)	(-1.90)
ln(P)		-0.0025***		-0.0025***	-0.0023***	-0.0013***	-0.0005***
		(-5.97)		(-5.89)	(-7.29)	(-6.17)	(-5.46)
Analysts		-0.0002***		-0.0002***	-0.0002***	-0.0001***	-0.0000***
		(-5.89)		(-5.89)	(-7.51)	(-6.43)	(-3.88)
Constant	0.0042***	0.0119***	0.0027***	0.0114***	0.0043**	0.0005	-0.0007
	(23.92)	(4.64)	(12.11)	(4.47)	(2.32)	(0.41)	(-1.11)
Observations	22737	18550	22737	18550	20200	20487	18751
R^2	0.026	0.059	0.032	0.060	0.053	0.041	0.022

t statistics in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table X

Analyst-Firm Level

The dependent variable is $\rho(3)_{i;j;t} = |FP(3)_{i;j;t} - FP(3)_{j;official;t}|/\sigma_{jt}(FP(3)_{i;j;t} - FP(3)_{j;official;t})$, which measures whether an analyst has the tendency to more strongly track the firm's official industry level forecast than all other analysts following the same firm j. *Experience* is the number of years the analyst has been issuing forecasts. *Firm Experience* is the number of years analyst i has been forecasting earnings for firm j. *Stocks* is the number of stocks for which analyst i has issued at least one forecast in year t. *Industries* is the number of different SIC2 industries covered by analyst i in year t. $Coverage_{i;j;t}$ is defined as $(\#\ Official_{i;j;t} - \#\ Fundamental_{i;j;t})/Stocks_{i;j;t}$ where $\#\ Official_{i;j;t}$ is the number of firms covered by analyst i in year t which belong to firm $j's$ official industry. $\#\ Fundamental_{i;j;t}$ is the number of Hoberg and Phillips peers of firm j covered by analyst i in year t. *Broker Size* is the number of analysts employed at analyst $i's$ brokerage house in year t. All regressions include the previously used firm-level control variables and year fixed effects. We restrict all regressions to forecasrs for which for which the stock price of the firm at the beginning of the quarter exceeds \$5 and for which the industry portfolios are populated by at least five different firms. Standard errors are clustered at the analyst level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\lvert FP(T)_{j;official;t} - FP(T)_{j;fundamental;t}\rvert$	51.5470***	43.0378***	44.8418***	39.8575***	-9.4689	54.2078***	53.3291***
	(14.01)	(7.95)	(10.34)	(6.35)	(-1.43)	(13.97)	(9.84)
$\lvert FP(T)_{j;official;t} - FP(T)_{j;fundamental;t}\rvert \times$ Experience$_{i;t}$		1.1799*					
		(1.95)					
$\lvert FP(T)_{j;official;t} - FP(T)_{j;fundamental;t}\rvert \times$ Firm Experience$_{i;j;t}$			1.8540**				
			(2.39)				
$\lvert FP(T)_{j;official;t} - FP(T)_{j;fundamental;t}\rvert \times$ Stocks$_{i;j;t}$				0.8468**			
				(2.18)			
$\lvert FP(T)_{j;official;t} - FP(T)_{j;fundamental;t}\rvert \times$ Industries$_{i;t}$					15.3842***		
					(10.35)		
$\lvert FP(T)_{j;official;t} - FP(T)_{j;fundamental;t}\rvert \times$ Coverage$_{i;j;t}$						-42.3956***	
						(-3.12)	
$\lvert FP(T)_{j;official;t} - FP(T)_{j;fundamental;t}\rvert \times$ Broker Size$_{i;t}$							-0.0384
							(-0.53)
Coverage$_{i;j;t}$						0.5621**	
						(2.05)	
Experience$_{i;t}$		-0.0043	0.0116	0.0114	0.0060	0.0093	0.0100
		(-0.19)	(0.57)	(0.56)	(0.30)	(0.44)	(0.49)
Firm Experience$_{i;j;t}$		-0.0789***	-0.1041***	-0.0792***	-0.0779***	-0.0772***	-0.0784***
		(-5.60)	(-6.27)	(-5.64)	(-5.54)	(-5.30)	(-5.55)
Stocks$_{i;t}$		0.0089	0.0087	-0.0000	0.0083	0.0087	0.0082
		(1.01)	(0.98)	(-0.00)	(0.92)	(0.93)	(0.91)
Industries$_{i;t}$		-0.0128	-0.0112	-0.0171	-0.2171***	-0.0069	0.0032
		(-0.33)	(-0.29)	(-0.44)	(-4.75)	(-0.16)	(0.08)
Broker Size$_{i;t}$		-0.0027	-0.0026	-0.0026	-0.0025	-0.0028	-0.0021
		(-1.47)	(-1.46)	(-1.45)	(-1.41)	(-1.50)	(-1.01)
Observations	114874	114874	114874	114874	113914	107183	113914
R^2	0.255	0.256	0.256	0.256	0.259	0.259	0.257
Analyst Fixed Effects	No	Yes	Yes	Yes	Yes	Yes	Yes

t statistics in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix

A De nition of return related variables

$SIC2\ firms$ is the number of di erent rms belonging to rm $j's$ two digit SIC industry in year $t-1$.

$Hoberg\ and\ Phillips\ firms$ is the number of di erent rms that are linked to rm j in the Hoberg and Phillips (2010a,b) sense in year t-1.

$ln(BE/ME)$ is the log of the rm's book to market equity. Book equity is de ned as ($at - lt - pstkl$ + $txditc$ + $dcvt$), market equity is calculated as $prc\_f \times csho$. In line with the asset pricing literature, the BE/ME ratio is updated at the end of each June in order to ensure a gap of at least 6 month between scal year ends and stock returns.

$Analysts_{j,t}$ is the number of di erent analysts having issued at least one scal year end EPS forecast in a given year. If there is no match between the CRSP/Compustat and the I/B/E/S datasets, it is assumed that the corresponding rm does not have any analyst coverage.

$r_{j,t}$ is the rm's monthly or weekly return from CRSP.

$r_{j,official,t}$ is the monthly or weekly return to an equally weighted portfolio consisting of all rms with the same SIC2 code.

$r_{j,fundamental,t}$ is the monthly or weekly return to an equally weighted portfolio consisting of all rms that are linked to rm j in the Hoberg and Phillips (2010a,b) sense. This variable varies at the rm-month level.

$\rho_{j,t}$ = $|(r_{j,t} - r_{j,official,t})/\sigma_{j,24}(r_{j,t} - r_{j,official,t})|$ is a measure that indicates to what extent the return of firm j has followed the return of its official SIC2 industry. $\sigma_{j,24}(r_{j,t} - r_{j,official,t})$ denotes the standard deviation of the return differential calculated for rolling windows of 24 months.

Low, Medium and High Official Industry Comovement indicate the first, second and third tercile of $\rho_{j,t-1}$ in each month.

$r_{j,t-2,t-12}$ is the firm's cumulative return between month $t-12$ and $t-2$.

$ln(M)_{j,t-1}$ is the natural log of the firm's market capitalization at the end of the previous month.

$Coverage_{i,j,t}$ is defined as $(\#\ Official_{i,j,t} - \#\ Fundamental_{i,j,t})/Stocks_{i,j,t}$.

Firm Experience$_{i,j,t}$ is the number of years analyst i has been forecasting earnings for firm j in year t.

B De nition of analyst related variables

B.1 Firm Level Variables

Main Variables

$ForecastError(T)_{j,t}$ is the average rm-level forecast error for all T horizon forecasts that were issued for rm $j's$ scal year end earnings in quarter t.

$FP(T)_{j,t}$ is the average T horizon forecast to price ratio at the rm level which is calculated by averaging across all outstanding T horizon forecasts for rm j in quarter t.

$FP(T)_{j,official,t}$ is the average T horizon forecast to price ratio prevailing in rm $j's$ o cial industry in quarter t. The o cial industry is de ned as the rm's two digit SIC industry. It is calculated using all outstanding forecasts for rms in the o cial sector. We require the o cial sector to be populated by at least ve di erent rms.

$FP(T)_{j,fundamental,t}$ is the average forecast to price ratio prevailing among rm $j's$ Hoberg and Phillips (2010a,b) siblings in the rst month of the quarter in which a forecast was issued. It is calculated using all outstanding forecasts for rms in the o cial sector. We require the rm to be linked to at least ve di erent rms in the Hoberg and Phillips (2010a,b) sense.

$\rho(T)_{j,t}$ is $|FP(T)_{j,t} - FP(T)_{j,official,t}|/\sigma_t(FP(T)_{j,t} - FP(T)_{j,official,t})$. σ_t is the cross sectional standard deviation in period t. The measure gives an indication about whether analysts as a whole tend to base their earnings forecasts for rm j on the average earnings forecast prevailing in rm $j's$ o cial industry category.

Control Variables

In all regressions of the second part of the paper we control for the following rm level characteristics.

$Analysts$ is the number of analysts with at least one year end EPS forecast throughout the current year.

Ret is the average monthly return over the previous calendar year.

$VolRet$ is the standard deviation of monthly stock returns over the previous calendar year.

$ln(Size)$ is the log of market capitalization at the end of the previous calendar year.

Roe is the return on equity in previous year $(oibdp-dp)/(at-lt-pstkl+txditc+dcvt)$.

$Volroe$ is the rm level standard deviation of return on equity between year $t-1$ and $t-5$.

$ln(BM)$ is the log of the rm's book to market equity at the end of the previous calendar year. Book equity is de ned as $(at-lt-pstkl+txditc+dcvt)$, market equity is calculated as $prc\_c \times csho$.

$SP500$ is a dummy indicating whether the company belongs to the SP500 index in the previous calendar year.

$ln(P)$ is the log of the rm's stock price at the beginning of quarter t.

B.2 Analyst and Analyst-Firm Level Variables

Analyst-Firm Level

The following variables are calculated at the forecast level.

$ForecastError(T)_{i,j,t}$ is the forecast error associated with the last T horizon forecast analyst i issues or revises in quarter t for year end earnings of firm j. The horizon T is measured in quarters.

$FP(3)_{i,j,t}$ is the forecast to price ratio in quarter t for analyst $i's$ three quarter ahead forecast of firm $j's$ fiscal year end earnings. It is calculated by normalizing the actual three quarter ahead forecast $F(3)_{i,j,t}$ by the stock price prevailing for firm j at the beginning of quarter t.

$\rho(3)_{i,j,t}$ is defined as $(|FP(3)_{i,j,t} - FP(3)_{j,official,t}|)/(\sigma_{jt}(FP(3)_{i,j,t} - FP(3)_{j,official,t})$

$\#\ Official_{i,j,t}$ is the number of SIC2 peers of company j covered by analyst i in year t.

$\#\ Official_{i,j,t}$ is the number of Hoberg and Phillips (2010a,b) peers of company j covered by analyst i in year t.

Analyst Level

The following variables are calculated at the analyst level.

$Stocks_{i,t}$ is the number of different stocks covered by analyst i in year t.

$Industries_{i,t}$ is the number of different industries (at the SIC2 level) covered by ana-

lyst i in year t.

$Experience_{i,t}$ is the number of years analyst i appears in the I/B/E/S database.

$BrokerSize_{i,t}$ is the number of analysts with at least one EPS forecast working at the same broker house as analyst i in year t.

Table A.I

Monthly Categorization Bias Strategy (EW)

At the beginning of each month, stocks are ranked according to the difference between an equally weighted portfolio of stocks belonging to the firm's official and an equally weighted portfolio consisting of stocks that belong to the firm's fundamental industry ($r_{j;official;t-1} - r_{j;fundamental;t-1}$). A firm's official industry is defined as the two digit SIC industry to which the firm belongs. A firm's fundamental industry is composed of all companies to which the firm has economic links in the Hoberg and Phillips (2010a,b) sense. The ranked stocks are assigned to one of five quintile portfolios. Q1 contains firms for which the return differential between the official and the fundamental industry is the lowest. Q1-Q5 is a portfolio that is long in stocks for which the return difference between the official and the fundamental industry portfolios has been the most negative (Q1, bottom 20%) and short in stocks for which the return differential has been the most positive (Q5, top 20%). All stocks are equally weighted within a quintile portfolio and rebalanced every month. Monthly alphas as well as Fama and French (1993) and Carhart (1997) factor loadings are in percent.

	alpha	mktrf	smb	hml	mom	R2	N
Q1	0.646***	0.871***	0.929***	0.0825	-0.0606	0.844	155
	(2.85)	(16.98)	(15.19)	(1.24)	(-1.56)		
Q2	0.140	0.891***	0.566***	0.249***	-0.0444**	0.922	155
	(1.10)	(30.89)	(16.46)	(6.66)	(-2.04)		
Q3	-0.189	0.884***	0.486***	0.430***	-0.0444*	0.878	155
	(-1.23)	(25.37)	(11.71)	(9.51)	(-1.69)		
Q4	-0.348**	0.977***	0.547***	0.321***	-0.0582**	0.905	155
	(-2.31)	(28.65)	(13.46)	(7.26)	(-2.26)		
Q5	-0.452**	1.083***	0.610***	0.194***	-0.107***	0.866	155
	(-2.15)	(22.78)	(10.76)	(3.15)	(-2.98)		
Q1-Q5	1.098***	-0.212**	0.319***	-0.112	0.0462	0.115	155
	(2.82)	(-2.40)	(3.03)	(-0.98)	(0.69)		

t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.II

Weekly Categorization Bias Strategy (EW): High Official Industry Mover

In this table we restrict the portfolios to *Strong Official Industry Follower* stocks (first tercile of ρ; see the appendix for a formal definition). We sort stocks that have abnormally closely followed their official industries in week $t-n$ (low ρ) into five quintiles of the difference between the stock's official and fundamental industry return during week $t-n$, i.e. $(r_{j;official;t-n} - r_{j;fundamental;t-n})$. All portfolios are equally weighted and rebalanced every week.

	alpha	mktrf	smb	hml	mom	R2	N
Panel A: Signal (t-1)							
Q1	0.308***	0.939***	0.847***	0.172***	-0.0922***	0.898	675
	(7.54)	(59.57)	(30.20)	(6.37)	(-5.64)		
Q5	-0.176***	1.000***	0.770***	0.0704**	-0.101***	0.903	675
	(-4.25)	(62.40)	(27.01)	(2.56)	(-6.07)		
Q1-Q5	0.484***	-0.0609**	0.0771*	0.102**	0.00863	0.0230	675
	(7.34)	(-2.39)	(1.70)	(2.33)	(0.33)		
Panel B: Signal (t-2)							
Q1	0.174***	0.939***	0.770***	0.194***	-0.0995***	0.892	674
	(4.20)	(58.74)	(27.08)	(7.08)	(-6.01)		
Q5	-0.0205	1.050***	0.791***	0.0677**	-0.0817***	0.913	674
	(-0.50)	(66.63)	(28.22)	(2.51)	(-5.01)		
Q1-Q5	0.194***	-0.110***	-0.0207	0.126***	-0.0178	0.0566	674
	(2.94)	(-4.33)	(-0.46)	(2.89)	(-0.67)		
Panel C: Signal (t-3)							
Q1	0.216***	0.996***	0.822***	0.190***	-0.0640***	0.905	673
	(5.33)	(63.60)	(29.52)	(7.10)	(-3.95)		
Q5	-0.0210	0.999***	0.691***	0.0312	-0.127***	0.895	673
	(-0.49)	(60.01)	(23.33)	(1.09)	(-7.34)		
Q1-Q5	0.237***	-0.00365	0.131***	0.159***	0.0627**	0.0302	673
	(3.46)	(-0.14)	(2.78)	(3.52)	(2.29)		
Panel D: Signal (t-4)							
Q1	0.134***	1.013***	0.795***	0.244***	-0.0621***	0.905	672
	(3.27)	(64.33)	(28.35)	(9.03)	(-3.81)		
Q5	0.00915	0.994***	0.744***	0.0856***	-0.133***	0.907	672
	(0.23)	(63.66)	(26.78)	(3.20)	(-8.25)		
Q1-Q5	0.124*	0.0195	0.0508	0.158***	0.0712***	0.0256	672
	(1.95)	(0.79)	(1.16)	(3.74)	(2.78)		
Panel E: Signal (t-5)							
Q1	0.0963**	1.001***	0.893***	0.119***	-0.0277*	0.912	671
	(2.42)	(65.21)	(32.70)	(4.54)	(-1.75)		
Q5	0.0630	0.977***	0.712***	0.104***	-0.184***	0.903	671
	(1.52)	(61.27)	(25.07)	(3.82)	(-11.16)		
Q1-Q5	0.0333	0.0233	0.181***	0.0148	0.157***	0.0951	671
	(0.53)	(0.97)	(4.22)	(0.36)	(6.26)		
Panel F: Signal (t-6)							
Q1	0.134***	1.042***	0.840***	0.100***	0.00481	0.913	670
	(3.33)	(67.23)	(30.40)	(3.79)	(0.30)		
Q5	0.0228	0.960***	0.680***	0.120***	-0.198***	0.893	670
	(0.53)	(58.12)	(23.09)	(4.24)	(-11.55)		
Q1-Q5	0.111*	0.0817***	0.160***	-0.0193	0.202***	0.133	670
	(1.72)	(3.28)	(3.60)	(-0.45)	(7.85)		

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.III

Fama-MacBeth (1973) Forecasting Regressions (Monthly Returns)

This table shows Fama and Macbeth forecasting (1973) regressions. The dependent variable is the monthly stock return of firm j. Equations (3), (4) and (5) are restricted to *Strong O cial Industry Follower* stocks. We restrict the regressions to stock-month observations for which the stock price at the end of the preceding month is greater or equal to $5. All variables are defined as in table VII).

	(1)	(2)	(3)	(4)	(5)
$r_{j;official;t-1} - r_{j;fundamental;t-1}$	-0.07 (-4.99)	-0.10 (-5.02)	-0.11 (-5.33)	-0.14 (-5.92)	-0.12 (-5.84)
Med Ind Follower$_{j;t-1}$		0.00 (1.90)			
Weak Ind Follower$_{j;t-1}$		0.00 (1.90)			
Med Ind Follower$_{j;t-1} \times (r_{j;official;t-1} - r_{j;fundamental;t-1})$		0.03 (2.58)			
Med Ind Follower$_{j;t-1} \times (r_{j;official;t-1} - r_{j;fundamental;t-1})$		0.04 (3.50)			
Mid Cap$_{j;t-1}$				0.00 (0.44)	
Large Cap$_{j;t-1}$				0.00 (0.49)	
Mid Cap$_{j;t-1} \times (r_{j;official;t-1} - r_{j;fundamental;t-1})$				0.03 (0.89)	
Large Cap$_{j;t-1} \times (r_{j;official;t-1} - r_{j;fundamental;t-1})$				0.06 (2.22)	
Med Analysts$_{j;t-1}$					0.00 (2.15)
High Analysts$_{j;t-1}$					0.01 (2.18)
Med Analyst$_{j;t-1} \times (r_{j;official;t-1} - r_{j;fundamental;t-1})$					0.01 (0.33)
High Analyst$_{j;t-1} \times (r_{official;t-1} - r_{fundamental;t-1})$					0.00 (-0.00)
$r_{j;t-1}$	-0.02 (-3.42)	-0.02 (-3.44)	0.05 (2.20)	0.05 (2.28)	0.05 (2.49)
$r_{j;t-2;t-12}$	0.00 (0.57)	0.00 (0.58)	0.00 (0.27)	0.00 (0.28)	0.00 (0.34)
ln(M)$_{j;t-1}$	0.00 (-0.45)	0.00 (-0.46)	0.00 (0.28)		
ln(BE/ME)	0.00 (0.51)	0.00 (0.50)	0.00 (0.95)	0.00 (0.94)	0.00 (1.23)

HAC adjusted t statistics in parentheses.