

# **An Experimental Investigation of the Positive and Negative Effects of Mutual Observation**

Robert Bloomfield  
Cornell University  
[rjb9@cornell.edu](mailto:rjb9@cornell.edu)  
(607) 255-9407

and

Jeffrey Hales  
The University of Texas at Austin  
[jeffrey.hales@mcombs.utexas.edu](mailto:jeffrey.hales@mcombs.utexas.edu)  
(512) 471-2163

First Draft: July 2005  
This Draft: May 2007

Helpful comments on this and prior versions of the paper were provided by Kendall Bowlin, Jason Brown, Michael Clement, Shana Clor, Harry Evans, Richard Frankel, Eric Hirst, Steve Kachelmeier, Ron King, Marlys Lipe, Lauren Maines, Don Moser, Mark Nelson, Kristiana Rennekamp, Bill Tayler, Jane Thayer, and workshop participants at Harvard University, the University of Pittsburgh, Washington University, the 2006 Lone Star Accounting Research Conference, the University of Texas Accounting and Finance Mini-Conference, and the 2006 AAA Annual Meeting. We also thank the Johnson Graduate School of Management and the McCombs School of Business for financial support.

# **An Experimental Investigation of the Positive and Negative Effects of Mutual Observation**

## **Abstract**

Despite the public nature of analyst forecasts, experimental research in accounting has largely ignored how the opportunity of forecasters to observe each other's opinions might influence the forecasts they produce. In this paper, we use two experiments to test predictions about the positive and negative impacts of such "mutual observation" on properties of consensus forecasts and on the judgments of investors who observe those forecasts. We find that when incentives for accuracy are high, mutual observation allows information aggregation, which increases both the accuracy and extremity of the consensus. However, when incentives are low, mutual observation induces free-riding, which offsets the beneficial information-aggregation effects. In our second experiment, we find that investors anticipate these effects and adjust their own forecasts accordingly.

# **An Experimental Investigation of the Positive and Negative Effects of Mutual Observation**

## **I. Introduction**

Despite the public nature of analyst forecasts, experimental research in accounting has focused almost exclusively on forecasting in individual settings (e.g., Maines and Hand 1996; Hunton and McEwen 1997; Sedor 2002), largely ignoring how the opportunity of forecasters to observe each other's opinions might influence the forecasts they produce. In this paper, we use two experiments to test predictions about the positive and negative impacts that such "mutual observation" can have on properties of consensus forecasts as well as on the judgments of investors who observe those forecasts. The key advantage of mutual observation is that it allows analysts the opportunity to learn from one another. The disadvantages are that mutual observation may induce free-riding (Kargin 2003), poor calibration may cause analysts to learn from one another in a biased way (Einhorn and Hogarth 1975; Dawes 1979), and investors who make decisions on the basis of consensus forecasts may fail to account appropriately for the redundancy that mutual observation introduces (Maines 1990, 1996; Soll 1999).

In addition to extending the experimental literature on analyst forecasts, our study has important practical implications. Over the last several years, analysts have been widely criticized for paying too much attention to one another, instead of providing their own independent analyses of fundamental information (Opdyke and Asinof 2001; Glassman 2001; Dreman 2002; Lynch 2002; Hulbert 2002; Ritholtz 2004; Shiller 2005), but to date there is little direct evidence to support these claims. This is likely due to the

fact that archival studies are not well-suited to assess the validity of such accusations because they lack a control group in which mutual observation is impossible.

To understand how mutual observation affects properties of the consensus, we form groups of four forecasters in our first experiment, present them with a residual-income forecasting task, and then manipulate whether or not they are allowed to observe their group's consensus while generating their own forecast. As in standard residual-income models (e.g., Frankel and Lee 1998), and consistent with empirical data (Dechow, Hutton, and Sloan 1999), we tell our forecasters that each firm's return on equity reverts (with noise) to its cost of capital, causing residual income to be imperfectly persistent. The forecasters use past time-series data to assess whether the firm's reversion rate is high or low, and this assessment determines their forecast. In this task, assessments of persistence affect forecast accuracy and forecaster payoffs the most when the target firm's return on equity is far from their cost of capital (i.e., when performance is abnormal).

We derive our predictions for this setting by assuming that forecasters are effort-averse Bayesians who share a prior belief about the time-series properties of earnings, and update their beliefs after observing a common data set of past earnings. Under these assumptions, individual analysts who are able to extract more information from the data construct posterior beliefs that are both more extreme and more accurate than those who extract less information. On the one hand, we expect mutual observation to have some positive benefits because it allows information aggregation. When the consensus is publicly observable, we normatively expect forecasters to adjust toward it, reducing forecast dispersion. Because this convergence of opinion reflects information

aggregation, it should, in theory, result in a more accurate and extreme consensus because the more informed (and therefore more extreme) individual forecasters will adjust less than will those forecasters who are less-informed. However, mutual observation may also produce negative effects. First, the ability to observe the consensus can induce free-riding because the consensus provides an alternative data source, reducing forecasters' incentives to extract information from the time-series data. Second, deviations from Bayesian rationality—such as a miscalibration of confidence and accuracy or a failure to account for forecast redundancy—can cause forecasters to misinterpret the consensus. Both free-riding and non-Bayesian reasoning could undermine the benefits of information aggregation and cause mutual observation to harm consensus accuracy.

As expected, our data from Experiment 1 reveal both positive and negative effects of mutual observation, and these effects are largely consistent with economic theory. To begin with, we find that mutual observation reduces forecast dispersion and increases consensus forecast extremity, indicating that mutual observation does allow analysts to aggregate the information they are able to individually extract from the common data set. We also find that impact of mutual observation on consensus forecast accuracy depends on target firm performance. When the target firm's return on equity is far from their cost of capital, so that analysts' incentives for accuracy are very high, mutual observation improves forecast accuracy. When incentives are low, the negative effects of mutual observation more than offset the positive effects of aggregation, consistent with the negative effects being driven by free-riding, not by miscalibration or a failure to account for redundancy.

Next, we conduct a second experiment to assess how mutual observation affects the usefulness of the consensus to others. In this second experiment, we present a new group of participants (whom we call “investors”) with the forecasts generated in Experiment 1 and ask them to make the same assessments as in Experiment 1 with similar incentives for accuracy. We tell the investors (truthfully) whether or not the analysts they observe from Experiment 1 were allowed to observe the consensus, but we do not provide investors with the fundamental information that was available to the analysts. Standard economic theory predicts that investors will fully anticipate the effects of the analysts’ ability (or inability) to observe the consensus. However, prior experimental research finds that people often overestimate the information contained in redundant signals (Maines 1990, Soll 1999), suggesting that investors may overreact to the high-extremity and low-dispersion forecasts generated by analysts in the observation condition by treating those forecasts as if they were independent.

Results of Experiment 2 again provide strong support for the predictions of economic theory. First, we find that investors who know that observation was not permitted enter forecasts that are more extreme than the consensus they observe, as if they understand that the more extreme individual forecasts are also more informative. In contrast, investors who know that observation was permitted enter forecasts that are no more extreme than the consensus, which is appropriate given that the consensus in these cases has already weighted the extreme forecasts more heavily. We also find that our investors (like our analyst forecasters in Experiment 1) are more accurate when abnormal performance increases the incentive to be accurate. Overall, investors appear to behave

quite rationally and account quite well for the forecast extremity and redundancy that mutual observation generates.

Our results have implications for recent research on (and criticism of) the informativeness of analyst reports by providing the first direct evidence on the advantages and disadvantages that mutual observation can have for analysts and investors. Our work extends prior psychological research on group forecasting because we use a setting that allows comparison to a normative benchmark. The normative benchmark allows us to distinguish our research from research on group-think (Janis 1982) and attitude polarization (Myers and Lamm 1976; Isenberg 1986), which show similar effects of group interaction on dispersion and extremity of opinion but without being able to draw clear inferences about whether those effects are normative. We also extend prior research by combining the literature on incentives (e.g., Libby and Lipe 1992; Bonner and Sprinkle 2002) with the literature on analyst forecasts and mutual observation. Doing so allows us to determine whether negative effects associated with mutual observation are driven by miscalibration or free-riding.

Our results also have some important practical implications. Most remedies for “analyst hype” have focused on limiting analysts’ incentives to bias their forecasts (for example, by making analysts more independent from investment bankers). However, such interventions do not address the free-riding problem we observe in our experiments. Free-riding might be an even more important issue than bias because bias can be undone by investors who are aware of analysts’ incentives. In contrast, investors have little recourse if analyst free-riding limits the amount of information that analysts analyze and make publicly available.

The remainder of the paper is as follows: In Section II we derive directional hypotheses from economic theory and discuss how those hypotheses relate to extant research in group psychology and economics. We describe the details and results of Experiment 1 in Section III. We describe the details and results of Experiment 2 in Section IV, and Section V concludes.

## **II. Theory and Hypotheses**

### **Background**

Analysts often have the opportunity to observe one another's published reports before making revisions to their predictions. Although the ability of analysts to revise their own forecasts after observing the consensus could be either beneficial or harmful, analysts have been widely criticized for failing to provide their own independent analyses of fundamental information (Opdyke and Asinof 2001; Glassman 2001; Dreman 2002; Lynch 2002; Hulbert 2002; Ritholtz 2004; Shiller 2005). Consider the following quote:

“I think that this [the tech bubble] is probably the largest bubble that we've seen at any time. Ironically, in this period we have had the best-trained analysts and money managers ever. We also had the best information available which is often almost instantaneous. Yet we've also had the biggest bubble in history....I believe the market experts led the change that drove tech stocks to unsustainable heights.” (Dreman 2002).

Dreman goes on to say, “I think there's more here than just greed. I think there's also the fact that the analysts are also affected by psychology.” Many pundits have warned of equity analysts' tendency to have a pack mentality. As one writer for the financial press put it, “Be skeptical of the experts....if a few top analysts start buying a story, then practically every analyst buys the story” (Glassman 2001). Similarly, academic researchers have warned of “thought contagions” in financial markets that are spread by

central communicators, such as analysts and other members of the financial press, which can lead to “mass delusion” or “irrational exuberance” (Lynch 2002; Shiller 2005; Seybert and Bloomfield 2006). These researchers argue for the use of contagion and epidemic models to describe the information technology and social psychology of financial markets, rather than relying on more traditional models that assume perfect rationality and market efficiency.

To date, however, there is surprisingly little empirical evidence to support these claims about analysts. Real world settings often confound the opportunity for mutual observation with other factors (such as reputation concerns) that will exert their own influence on consensus accuracy, making it difficult to disentangle the underlying determinants of analyst behavior. Even if these confounds could be addressed through careful econometric analysis, archival studies lack a control group in which mutual observation among analysts is impossible, limiting the ability of archival studies to speak directly to the effect of mutual observation.

These data limitations could be overcome in experimental settings, but much of the experimental research in accounting has examined forecaster behavior in individual settings. For example, researchers have investigated whether forecasters understand the time-series properties of earnings (Maines and Hand 1996), whether forecasters are affected by the form and format of disclosures (Maines, McDaniel, and Harris 1997; Sedor 2002; Krische 2005), and how forecasters respond to motivational incentives (Hunton and McEwen 1997). While these studies provide important insights into how various factors may influence individual forecasters, we know of no experimental study on earnings forecasts that has directly manipulated mutual observation.

Empirical study of mutual observation is important because it offers the potential for both positive and negative effects. On the one hand, we expect mutual observation to have positive effects on forecast accuracy because limits to skill and effort lead individual analysts to capture incompletely the implications of a data set (as in Bloomfield 2002). Mutual observation is then a mechanism through which analysts can aggregate their information by placing more weight on the forecasts of those who (through greater effort or skill) draw more information from the available data, resulting in a consensus that is more accurate than a simple weighted average of their independent efforts – a benefit of mutual observation rarely emphasized when analysts are being criticized for having a pack mentality. On the other hand, we expect these positive effects to be offset because mutual observation induces analysts to exert less effort in individually analyzing data given that any one analyst could still achieve an accurate forecast by learning from the others. Such free-riding reduces the total information that is available to aggregate, and this effect will be particularly strong if low incentives make free-riding relatively attractive (Kargin 2003).

In addition to the economic motivation for free-riding, mutual observation could also have negative effects that are more psychologically based. When analysts' confidence levels are not calibrated to match the true quality of their information, mutual observation can result in a consensus that places too much weight on inaccurate forecasts and too little weight on accurate forecasts (Einhorn and Hogarth 1975; Dawes 1979). In addition, a failure to account for redundancy can lead analysts and investors to become too confident in the consensus (Maines 1990, 1996).

## Research Setting

To examine the advantages and disadvantages of mutual observation in an objective setting, we present participants with an earnings forecasting task within a residual-income valuation framework (Frankel and Lee 1998; Dechow et al. 1999). Such models determine value-to-book ratios using projections of abnormal earnings, defined as return-on-equity (ROE) less the cost of capital. We tell forecasters that abnormal earnings revert to a cost of capital of 10%, with the expected abnormal earnings being a fraction,  $\rho$ , of the prior year's abnormal earnings. Forecasters are also told that this fraction is constant for each firm, and is equally likely to be 95% (high persistence) or 65% (low persistence). Thus,  $E[\text{ROE}_{t+1}|\text{ROE}_t] = 0.65(\text{ROE}_t - 10\%) + 0.35(10\%)$  when persistence is low, and  $E[\text{ROE}_{t+1}|\text{ROE}_t] = 0.95(\text{ROE}_t - 10\%) + 0.05(10\%)$  when persistence is high. Finally, forecasters are told that the standard deviation around this expectation is 4%. Forecasters are shown six or seven consecutive periods of data, and are asked to assess the probability (from 0% to 100%) that the firm's abnormal earnings has high persistence.

A rational Bayesian who observes two consecutive earnings realizations in this setting computes the probability of high persistence by computing the probability of observing the second earnings number under high and low persistence. As an example, suppose that ROE in period 1 was 15%. If ROE in period 2 were 12%, standard Bayesian reasoning gives the following for  $\alpha$ :

$$\begin{aligned}\alpha = \Pr(\rho = 0.95 \mid r_1, r_2) &= \frac{\Pr(r_1=0.15 \text{ and } r_2=0.12 \mid \rho=0.95)}{\Pr(r_1=0.15 \text{ and } r_2=0.12 \mid \rho=0.95) + \Pr(r_1=0.15 \text{ and } r_2=0.12 \mid \rho=0.65)} \\ &= 0.315 / (0.315 + 0.380) = 45.3\%\end{aligned}$$

Each additional period provides an independent observation. For  $j$  observations,

$$\alpha = \frac{\Pr(r_1, r_2, \dots, r_j \mid \rho=0.95)}{\Pr(r_1, r_2, \dots, r_j \mid \rho=0.95) + \Pr(r_1, r_2, \dots, r_j \mid \rho=0.65)} \quad (1)$$

As long as each data point has some (but not perfectly diagnostic) information, obtaining more observations will increase the expected diagnosticity of the analysis, and thus increase the expected extremity of  $\alpha$  relative to the shared prior of  $\Pr(\rho = 0.95 \mid \emptyset) = 0.5$ .

We derive our predictions from three behaviorally motivated assumptions. First, we assume that limited information processing capacity causes each individual analyst to capture the implications of the data set incompletely (as in Bloomfield 2002). More formally, each analyst,  $j$ , can be thought of as selecting a random sample of data points to observe and each of those true data points,  $r_i$ , gets encoded as  $\hat{r}_{i,j} = r_i + z_{i,j}$ , where  $z_{i,j}$  is an idiosyncratic, noisy (but unbiased) normally-distributed error term with variance  $s_{i,j}$ . This assumption allows us to use traditional Bayesian reasoning, while recognizing that analysts are imperfect information processors.

Second, we assume that mutual observation is at least partly effective in allowing analysts to aggregate their information. The noise terms used to reflect each analyst's imperfect processing are independent of one another, so that analysts hold more information collectively than is reflected in any one of their individual forecasts. Analysts' limited information processing capacity is likely to make aggregation incomplete, just as it makes data extraction incomplete. We do not model this incompleteness, but simply assume that mutual observation allows some degree of aggregation.

Third, we assume that analysts are effort averse, so that they will observe fewer points, and/or observe those points with greater noise, when their incentives for accuracy

are lower. In particular, the more information forecasters have already extracted from the data, the less valuable additional information extraction will be. Because mutual observation provides an alternative source of information (the consensus), it necessarily reduces the incentive for any one individual who cares about their forecast accuracy to extract additional information.<sup>1</sup> If information extraction were trivial, effort aversion might have little impact. We expect to see evidence of this effect, however, because information extraction is quite time-consuming and challenging in our setting: to analyze the implications of a single data point, one must compute the expected ROE under high and low persistence, calculate the distance of realized ROE from each expectation, calculate the probabilities associated with each distance, and then incorporate those probabilities into equation (1). This aspect of our task is important because, as we will see, it allows us to distinguish the influence of free-riding from other behaviors predicted by psychology.

Our three assumptions generate two hypotheses. Our first hypothesis addresses the effect of mutual observation on forecast dispersion. Because it is at least partly effective in allowing analysts to aggregate their information, mutual observation will increase the similarity of the analysts' estimates (by allowing them to share, and thereby cancel out, their independent noise terms). As a result, we predict that mutual observation will reduce forecast dispersion around the consensus:

### **H1. Mutual observation lowers within-group forecast dispersion.**

---

<sup>1</sup> For further discussion of this topic, see Kargin (2003), who derives a unique compensation scheme specifically to avoid this incentive problem.

Our second hypothesis addresses the effect of mutual observation on the accuracy and extremity of the consensus forecast. Mutual observation has two offsetting effects on accuracy and extremity. Information aggregation increases both accuracy and extremity of the forecast by placing greater weight on more extreme and (in expectation) more accurate forecasts. Recall that, among rational Bayesians, accuracy and extremity vary together in our RIM forecasting context. Rational aggregation, even if imperfect, will place greater weight on the more accurate and extreme forecasts, and therefore result in consensus estimates that are also more accurate and extreme. (A natural aggregation mechanism would be that those who are more informed and extreme adjust less toward the consensus than those who are less informed and extreme).

The effects of mutual observation on aggregation are offset at least partly by the effect of free-riding. As discussed above, mutual observation reduces information extraction by providing an alternative source of information (the consensus). The reduction in information extracted leads directly to a reduction in the accuracy and extremity of the consensus forecast. Because aggregation and free-riding have offsetting effects on accuracy and extremity, we cannot hypothesize a main effect of mutual observation on these variables. However, mutual observation will be less likely to cause free-riding when the firm experiences abnormal performance because assessments of persistence for these firms have much more impact on forecasters' payoffs. As a result, we predict that extremity and accuracy will be influenced by an interaction between observation and firm performance.

**H2. Mutual observation increases the extremity and improves the accuracy of the consensus the most for firms with abnormal performance.**

We have derived H1 and H2 primarily from economic assumptions about Bayesian updating, information aggregation, and effort aversion, with our primary deviation from traditional economic models being that we assume information extraction is imperfect. Research in psychology supports many, but not all, of our predictions. Studies on “groupthink” (Janis 1982) and attitude polarization (Myers and Lamm 1976; Isenberg 1986) show that individuals with moderate opinions become more extreme when they interact in small groups. However, most studies in psychology argue that such results effects are non-normative, either because participants are most persuaded by those with the most extreme beliefs, rather than those with the most expertise or information (a miscalibration effect) or because people infer that agreement among many people indicates independent confirmation even though the judgments are not independent (a redundancy effect).

Our experiment contributes to this psychological literature in part because two features of the RIM forecasting task differ significantly from most prior group judgment tasks. Most psychological studies examine groupthink and polarization by asking participants to form opinions about political issues (such as the desirability of the death penalty or abortion rights) or to answer almanac questions (like “How long is the Nile River?”). In these settings, it is difficult to assess whether group interaction leads to “too much” extremity. Political opinions cannot be compared to a normative benchmark, and almanac questions, while having objective answers, are rarely chosen randomly. In such cases, it is difficult to determine whether systematic errors represent biased judgment or a misunderstanding of how questions are selected (Gigerenzer, Hoffrage, and Kleinbolting 1991). In contrast, by providing participants with objective information in a setting with

an optimal answer, we are able to assess whether group observation causes the consensus forecast to become too extreme.

Another important difference is that, unlike prior psychological research, participants in our setting share a prior belief about the appropriate forecast, and deviate from that prior belief in response to data. The shared prior facilitates information aggregation by allowing forecasters to signal their expected confidence and accuracy (via forecast extremity). Without a shared prior, it would be more difficult for analysts and investors to aggregate information. In our setting, as with earnings forecasting more generally, analysts whose forecasts deviate more from the baseline prior presumably deviate because they have information (rather than a lack of information) warranting such an extreme deviation. This may explain why recent research has found that bold forecasts (i.e., those that deviate more from the consensus) tend to be more informative than herding forecasts (Clement and Tse 2005).

Lastly, our setting allows for a natural manipulation of analysts' incentives for accuracy (as explained previously) by allowing target firm performance to vary. In doing so, we integrate the literature on incentives (e.g., Libby and Lipe 1992; Bonner and Sprinkle 2002) with the literature on analyst forecasts and mutual observation. This feature of our experimental design is important because it allows us to distinguish whether negative effects (when and if they appear) are driven by free-riding or non-Bayesian reasoning. If negative effects are driven by free-riding, we should see their influence vary with incentives (as we hypothesize in H2).

### III. Experiment 1

#### Method

##### *Design Overview*

The experiment involved 68 MBA students at a top-ranked business school. Participants made two forecasts for each of eight fictitious companies (16 total forecasts) using a residual income model (RIM). Participants were randomly assigned to 17 groups of four participants each. Using a computerized network, ten of the groups were allowed within-group mutual observation, as we explain below, while the remaining seven groups were forced to make independent decisions without observing other participants' forecasts. Within both settings, half of the groups saw the eight firms in one order, and the other half saw the firms in the opposite order. The experiment, therefore, uses a 2 (Observation) \* 2 (Order) \* 16 (Period) factorial design. The first two factors are manipulated between groups. The last factor, which is manipulated within groups, allows us to assess the impact of abnormal performance.

##### *The Task*

All participants had received previous RIM training, either in a course on financial statement analysis or in a course on topics in financial markets. Instructions were distributed before each session and reviewed by the experiment administrator before participants made any decisions. The instructions explained that participants would be assessing the time-series process governing earnings twice for each of eight fictitious firms (see Table 1). Performance for each firm was created by a random number generator, with expected return on equity (ROE) in one year being a weighted average of

the prior year's ROE and the cost of capital, which was set to 10%. The standard deviation of ROE around this expectation was 4%. Participants were also told that for any given firm there was a 50% chance that the weighting on prior ROE would be 95% (high persistence) and a 50% chance that the weighting on prior ROE would be 65% (low persistence).

For each firm, participants initially saw a screen that included a tabular display of the first six years of earnings, book value, and ROE (Figure 1, Panel A). The screen also included a graph of ROE for those years (Figure 1, Panel B). Based on this data, participants were asked to assess the probability that the firm's earnings were being generated with high persistence. This assessment is  $\alpha$  from equation (1). Participants submitted their assessment of  $\alpha$  by choosing a number between 0 and 100 from a drop-down menu. Given this probability assessment, the software automatically projected for the participant next period earnings and ROE, as well as an appropriate current value-to-book ratio using the standard methods of residual-income valuation. Each of these projections was automatically displayed on the computer screen and updated immediately for any revision in the participant's assessment of  $\alpha$ . Participants were allowed to revise their assessment (called their "tentative" assessment) as many times as they wanted, until they had clicked a button to transmit their "final" assessment; at that point, they could not change their assessment again.

After all four participants in a group had submitted their final individual assessment of  $\alpha$ , actual earnings and ROE were then revealed for the year just forecasted. Participants then made another assessment of  $\alpha$ , now based on the original time series plus one additional year of data. Upon completion of this second assessment, actual

earnings and ROE for this second forecasted year were revealed, as which point the firm's true persistence level was also revealed along with firm value for both periods and how much the participants earned in those periods for their estimates (as we describe below). Participants then continued on to the next firm, repeating the same steps for each firm. Note that at no time were participants told what the optimal assessment of  $\alpha$  was given the data that they saw, only whether high or low persistence had actually been used in generating a given firm's earnings.

### ***Manipulating Observation***

In our observation setting, participants were always able to see their group's current consensus estimate of  $\alpha$ , computed as the mean of their individual assessments. Participants could see how many other (but not which) group members had submitted an estimate for inclusion in the consensus. When a participant submitted a tentative assessment or a revision, the consensus was updated to reflect this new information and the update was immediately relayed for all group members to see. Although they could not directly revoke an assessment, participants could (as described above) revise their tentative assessments freely. The consensus always reflected, and only reflected, the most current information available from each person in the group. In our no-observation setting, participants received no information about the assessments of their fellow group members, but their task was identical in all other respects.<sup>2</sup>

---

<sup>2</sup> Our instructions allow participants to infer that we manipulate observation between cohorts, perhaps leading participants in the observation setting to attend more or less to the consensus, depending on the inference (if any) they drew from the manipulation. Either way, attending to the consensus for non-informational reasons would seem to bias against supporting our Bayesian-based predictions.

## *Incentives*

Participants' incentives are based on the accuracy of the value estimates derived from their probability estimates. We use a version of the Becker, DeGroot, and Marschak (1964) device to ensure incentive compatibility. We first compute the "baseline" value of a firm with a 50% probability of high persistence. If the participant's value estimate is greater (less) than the baseline value, the participant buys (sells) one share at every price from the baseline value to their estimated value. Gains or losses on these trades were determined by the difference between the trade price and the true value (given the true level of persistence). This payment scheme is incentive compatible because participants maximize their expected profit by accurately revealing their estimate; doing otherwise forces the participants to engage in trades they believe will be unprofitable or to miss an opportunity to engage in some profitable trades.

Participants in both the observation and no-observation settings received 10% of the gain or loss of the other participants in their group. We included this feature in the observation setting to ensure that participants did not believe they had an incentive to withhold information from the group by delaying the submission of a tentative estimate or an incentive to trick others in their group by entering misleading tentative assessments. We included this feature in the no-observation setting to avoid confounding the effects of observational learning with a difference in the incentive scheme.

Gains and losses in laboratory dollars were converted to US dollars by adding a constant to their laboratory dollar amount, and then converting with an exchange rate. Participants were paid either the calculated amount or \$5, whichever was greater. Neither the constant nor the exchange rate was revealed to participants to reduce the

chances of risk-seeking behavior of participants who suspected they were close to receiving the minimum payment.

### *Eliminating Structural Concerns*

Prior research has offered two alternative economic reasons for why mutual observation might fail to aggregate all available information into the consensus: informational cascades and rational herding. In an informational cascade, public information swamps private information so that analysts rationally discard their private information and always agree with the consensus (Welch 1992; Bikhchandani, Hirshleifer, and Welch 1998). With rational herding, analysts have economic incentives to shade their forecasts away from their private individual assessments so that their public forecasts are closer to the consensus (e.g., Scharfstein and Stein 1990; Trueman 1994; Zwiebel 1995; Hong, Kubik, and Salomon 2000).

Informational cascades and rational herding would make it difficult to detect other influences on analyst behavior, so we carefully designed our experiment to avoid both. We avoid informational cascades by allowing the analysts great flexibility in adjusting their forecasts. Each analyst can enter their first forecast whenever they wish and can revise their forecast whenever and as often as they wish. We avoid rational herding because, even within their groups, the analysts remain anonymous, and they are given incentives for absolute, not relative, accuracy (e.g., Francis and Philbrick 1993; Jegadeesh and Kim 2006).

## Results

Our hypotheses predict the effects of observation and abnormal performance on the dispersion, extremity, and accuracy of consensus forecasts. For simplicity of presentation and statistical analysis, we rank the 16 events according to the absolute magnitude of residual income, and classify them into two levels of performance: normal (the absolute difference between ROE and cost of capital is in the lowest three quartiles of performance) or abnormal (the absolute difference between ROE and cost of capital is in largest quartile).<sup>3</sup> For all statistical tests, we also include a variable that accounts for the diagnosticity of the time-series evidence. The more diagnostic the evidence, the more extreme is the optimal assessment of persistence. We control for diagnosticity because forecast errors are likely to be correlated with diagnosticity (Bloomfield, Libby, and Nelson 2000) and because firms with performance far from the cost of capital are likely to have high persistence. As with abnormal performance, we classify the 16 events into two levels of diagnosticity: low (the absolute difference between the probability of high persistence and 50% is in the lowest three quartiles) or high (the absolute difference is in the highest quartile).

Forecast dispersion as a function of observation and performance is shown in Panel A of Figure 2. We define forecast dispersion for each group of forecasters as the sum of the absolute distance of the forecasts from the consensus. As predicted, mean dispersion is substantially lower when observation is allowed, averaging 12.10 over all events with observation, compared to 21.93 without observation. To test the significance

---

<sup>3</sup> Most analyses show little difference among events within the bottom three quartiles, while there are substantial differences between these securities and those in the top quartile. This pattern is consistent with sensitivity of financial performance to persistence assessments in the two groups. Our inferences are similar when the analysis is done using quartiles, rather than the binary classification.

of this difference (and for all other analyses), we use a repeated-measures ANOVA, with between-group factors for observation and firm order and within-group factors for abnormal performance and diagnosticity. We exclude any factor representing firm or year, as these were rarely significant either separately or in interaction with other variables, and have little effect on the inferences of interest. We find that observation has a statistically significant effect on dispersion ( $p < 0.001$ , one-tailed), providing clear evidence that forecasters are responding to one another and converging on an outcome. Thus, we provide strong support for H1. We also note support for H1 is strongest when performance is abnormal ( $p < 0.001$ ). This suggests that observation may result in more learning when performance is abnormal, perhaps because forecasters devote greater effort to observation as their incentives increase.

Panel B of Figure 2 plots consensus extremity as a function of observation and performance. Averaging over all events, observation increases extremity from 14.96 to 21.45, with the main effect of observation statistically significant ( $p = 0.004$ , one-tailed). Consistent with H2, we also observe a significant interaction between observation and performance ( $p = 0.008$ ). While observation has a relatively small effect for ordinary performance ( $\mu_{\text{no obs}} = 13.78$  vs.  $\mu_{\text{obs}} = 18.45$ ;  $p = 0.053$ , one-tailed), it greatly increases extremity for abnormal performance ( $\mu_{\text{no obs}} = 18.48$  vs.  $\mu_{\text{obs}} = 30.44$ ;  $p = 0.002$ , one-tailed). We find little evidence that miscalibration leads to excess extremity. On the contrary, when performance is normal, the consensus is only nominally higher than optimal extremity ( $\mu_{\text{difference}} = 1.64$ ,  $p = 0.231$ ), and the consensus is insufficiently extreme, rather than too extreme, when performance is abnormal ( $\mu_{\text{difference}} = -5.89$ ,  $p = 0.026$ ).

Looking at the absolute error of the consensus, Panel C of Figure 2 shows an interaction between observation and performance, which provides additional support for H2: observation is far more beneficial when performance is abnormal than when performance is normal ( $p = 0.005$  for interaction). Analysis of the simple effects indicates that observation harms the consensus when performance is normal ( $p = 0.044$ , one-tailed) but improves the consensus when performance is abnormal ( $p = 0.095$ , one-tailed). These results do provide some evidence that observation is harmful, consistent with behaviorally-induced concerns about observation, but only when incentives are low.

Overall, the results from Experiment 1 indicate that mutual observation will affect the informativeness of the consensus through the countervailing effects of Bayesian updating and free-riding, with the latter being suppressed the most when incentives for accuracy are the highest. Supplementary analyses reinforce this interpretation. For example, an untabulated analysis of individual forecast accuracy shows that observation greatly reduces individual forecast errors when performance is abnormal ( $\mu_{\text{no obs}} = 28.73$  vs.  $\mu_{\text{obs}} = 19.39$ ;  $p = 0.011$ ), but not when performance is normal ( $\mu_{\text{no obs}} = 27.59$  vs.  $\mu_{\text{obs}} = 25.53$ ;  $p = 0.659$ ), consistent with the economic theory that the ability to observe others should not make an individual worse off, and may help.<sup>4</sup> The significant observation-performance interaction suggests that, when performance is normal (so that incentives are moderate), forecasters use observation as a *substitute* for individual effort, achieving similar levels of individual accuracy with less effort than if they worked independently; when firm performance is abnormal (so that incentives are strong), forecasters use

---

<sup>4</sup> The insignificant main effect of observation when firm performance is normal is qualified by a significant interaction between observation and diagnosticity ( $p = 0.030$ ): at low levels of diagnosticity, observation lowers the magnitude of individual forecast errors, but increases forecast error when diagnosticity is high.

observation as a *supplement* to effort, dramatically improving their individual performance.

#### **IV. Experiment 2**

The results of Experiment 1 show that mutual observation dramatically changes the dispersion, extremity, and accuracy of analyst forecasts. We now examine whether investors account for these changes when they base their own forecasts on analysts' reports. In our second experiment, we present the forecasts generated in Experiment 1 to a group of student participants we call "investors." The investors, who are given incentives to be accurate, do not see the time-series information presented to the analysts. Our key manipulation is whether investors are presented with the assessments of analysts who were in our observation setting or in our no-observation setting.

#### **Hypotheses**

As in Experiment 1, we begin by deriving our predictions from economic theory. When analyst forecasts were created without the possibility of mutual observation, investors can use the extremity of each forecast as an indication of its statistical precision. Investors can use the precision-extremity relationship to conduct the same type of aggregation that would have occurred through mutual observation, placing more weight on more extreme forecasts. Consequently, we expect investor forecasts to be more extreme and more accurate than the consensus forecasts they observe, because the consensus weights all four independent forecasts equally. In contrast, when investors observe analyst forecasts that were generated in the presence of mutual observation, the

consensus they see already reflects greater weight on the opinions of those analysts who were able to extract the most information out of the data. Therefore, investors in the observation condition can outperform the consensus only if they are able to *incrementally* improve upon the information aggregation that the analysts did themselves. Because we see no reason for investors to be incrementally superior to analysts, we make the following prediction:

**H3. Investor forecasts are more extreme and accurate than the consensus they observe in the no-observation setting, but not in the observation setting.**

Because investors can aggregate independent forecasts themselves, mutual observation offers no positive information-aggregation effect for investors. On the contrary, it may create a disadvantage because investors cannot recreate information lost due to free-riding (because they do not have access to the data that analysts saw). As a result, investors who see independent forecasts have a theoretical information advantage over investors in the observation condition, the magnitude of which will be greatest when firm performance is normal (because low incentives encourage free-riding), as stated in our final prediction:

**H4. Investor forecasts are more extreme and accurate in the no-observation setting than in the observation setting, particularly when firm performance is normal.**

Psychological theory suggests some clear alternatives to H3 and H4. Maines (1990) found that individuals ignore the implications of forecast redundancy when assessing the expected accuracy of a consensus forecast, and Maines (1996) found that

manipulating forecaster dependence did not alter how participants combined analyst forecasts together. If investors ignore forecast redundancy and treat analysts' forecasts in the observation setting as if they were produced independently, then in both conditions we will see investor forecasts that are more extreme than the consensus. This failure to account for redundancy would also lead investor forecasts to be more extreme, but less accurate, in the observation setting than the no-observation setting.

Another alternative hypothesis is suggested by Larrick and Soll (2006) and Soll and Larrick (2006). These papers show that people have poor intuitions about how to combine the opinions of others or to revise their own opinion in lights of the opinions of others. In particular, people often choose from among a set of experts, rather than averaging the experts' opinions. Untabulated data from Experiment 1 indicates that *individual* analyst forecasts in the no-observation tend to be more extreme but less accurate than individual analyst forecasts in the observation setting. As a result, if investors in both of our settings simply choose to agree with a single analyst for any given firm, investors in the no-observation setting would tend to be more extreme but less accurate than investors in the observation setting.

## **Method**

### ***Experimental Design***

The experiment involved 34 MBA students at a top-ranked public business school. Participants were randomly assigned to one of two settings. Twenty individuals were assigned to observe the assessments produced in the observation setting of Experiment 1, and 14 individuals were assigned to observe the assessments produced in

the no-observation setting of Experiment 1. Within each condition, we “yoked” two investor participants to each of the Experiment 1 analyst cohorts. Because firm order had little influence in Experiment 1, we used the same firm order for all participants in Experiment 2. The experiment uses a 2 (Observation Setting) \* N (Analyst Cohort: 10 or 7) \* 16 (Period) nested factorial design. The first two factors are manipulated between participants. The last factor is a within-participant repeated measure.

### ***The Task***

For each of the 16 periods, investors were asked to evaluate the assessments of the four analysts to whom they were yoked. In addition to seeing the analysts’ assessments of  $\alpha$  (as shown in Figure 3), we also provided investors with the payoff implications for agreeing with each analyst’s assessment (both in tabular and graphical format). We did not provide them with the fundamental information observed by the analysts. After reviewing this information for a given firm-period, investors were asked to make their own assessment of  $\alpha$ . Unlike Experiment 1, our investor participants in Experiment 2 did not learn actual persistence for any of the firms.

### ***Manipulation of Analyst Setting***

Because investors saw assessments that were generated either with or without mutual observation, they saw analyst forecasts that, on average, had very different properties (i.e., dispersion, extremity, and accuracy). In addition to the manipulation of what type of forecasts investors saw, we also disclosed the context in which the assessments were made. Investors in the no-observation setting were simply told the following: “The four participants you will see from Part 1 were not allowed to

communicate with one another and so had no knowledge of each other's judgments." In contrast, investors in the observation setting were given the following information:

"The four participants you will see from Part 1 were allowed to see one another's judgments. Specifically, while making their judgments, these participants could see the current average of the judgments submitted by their group members, and could adjust their own judgments as often as they wished before submitting a final judgment. If a participant updated his or her judgment, the average was recalculated using the new judgment and the average was immediately communicated to all four participants in the group. All of the judgments you will see reflect the final judgments submitted by these four participants for each firm."

All investor participants were told that the analyst forecasters from Experiment 1 were paid according to the accuracy of the value estimate that followed from their judgment.

### *Incentives*

Our investor participants faced the same basic incentives as the analysts in Experiment 1 with two exceptions: analyst participants in Experiment 1 were also compensated for the accuracy of the consensus and investor participants in Experiment 2 had the opportunity to scale up or down the payoff implications of their individual assessments.<sup>5</sup> Because use of the multiplier did not differ across treatments and does not alter our inferences, we do not discuss it further.

Upon completion of the experiment, gains and losses in laboratory dollars were converted to US dollars by adding a constant to their laboratory dollar amount, and then converting it with an exchange rate. Participants were paid either the calculated amount

---

<sup>5</sup> After making their own assessment, investors were asked to choose a multiplier (1/4, 1/2, 1/3, 1, 2, 3, or 4). Choosing a multiplier of one resulted in a payment (in laboratory dollars) that was exactly equal to the corresponding payoff implications as displayed on the graph that they saw (which would be identical to what an analyst would have received in individual compensation if they had made the same assessment). Choosing a multiplier of less than (greater than) one reduced (increased) the gain or loss relative to what it would have been with a multiplier of one. Investors, therefore, maximized their expected gain by accurately reporting their probability judgment and choosing the largest multiplier.

or \$5, whichever was greater. Neither the constant nor the exchange rate was revealed to participants. Average compensation was \$15.

## Results

Figure 4 displays the extremity of investors' assessments of persistence conditional on firm performance. We code performance exactly as in the analysis of Experiment 1. We replace the factor for diagnosticity of the time-series information (which investors did not have access to) with a factor to control for the extremity of the consensus forecast. We construct this control factor by ranking consensus estimates by extremity and classifying the consensus as moderate (if in the lowest three quartiles of extremity) or extreme (if in the highest quartile of extremity).<sup>6</sup>

As shown in Panel A of Figure 4, investor extremity, relative to the analyst consensus, depends strongly on analysts' opportunity for mutual observation ( $p = 0.010$ , one-tailed), as predicted in H3. In the no-observation setting, investors are significantly more extreme than the analyst consensus ( $\mu = 8.49$ ,  $p < 0.001$ ). In the observation setting, investor forecasts are nominally *less* extreme than the analyst consensus ( $\mu = -1.22$ ,  $p = 0.396$ ). Thus, our results confirm H3, with respect to extremity.

To further clarify how investors react to the consensus and to provide additional support for the theory underlying H3, we estimated a regression model with investor extremity as the dependent variable and consensus extremity and dispersion as independent variables. The Bayesian model predicts that investor forecasts should be more extreme when the consensus is more extreme. In addition, investor extremity should be *increasing* in consensus dispersion in the no observation setting, but not in the

---

<sup>6</sup> Inferences are similar when the analysis is done using quartiles rather than a binary classification.

observation setting. To see why, recall that Bayesian aggregation places greater weight on more extreme forecasts. Therefore, the only time those investors should be no more extreme than the consensus is when forecast dispersion is zero. Because investors in the observation setting do not need to aggregate forecasts, we expect investors to weight all forecasts equally and so to be unaffected by the degree of dispersion.

Regression results strongly support our Bayesian predictions. Coefficients on dispersion are positive in the no-observation setting (0.34,  $p < 0.010$ , one-tailed), while being indistinguishable from 0 in the observation setting (-0.03,  $p = 0.676$ ). The difference in the coefficients is statistically significant ( $p = 0.006$ , one-tailed). In these regressions, we also find that the coefficients on consensus extremity tend to be positive ( $p < 0.001$ ) and are almost identical in the two settings (0.74 in the no-observation setting and 0.75 in the observation setting).<sup>7</sup>

The data provide little support for our predictions about accuracy. We observe no main effect of observation, or interactions of observation with other variables, when analyzing relative or absolute levels of accuracy. One possible explanation for these non-results is that we simply lack the power to observe these effects. The accuracy effects in Experiment 1 (as shown in Panel C of Figure 2) are relatively small, so responses to those differences are hard to detect. Moreover, the expected results could also have been swamped by an effect we did not predict: investors in the no-observation setting do a worse job of combining analyst opinions when incentives are low ( $p = 0.050$ ). When target firm performance is abnormal, investors in the no-observation setting do nominally better than the consensus, but they do much worse than the consensus when performance

---

<sup>7</sup> Rather than pooling all 544 investor-year observations into a signal regression, we fit a separate regression for each investor's 16 individual forecasts and then analyze the coefficients from these 34 regressions to avoid overstating the statistical significance of the regression results.

is normal. As a consequence, we find an overall main effect of performance on absolute forecast error ( $p = 0.018$ ), rather than the expected interaction with observation. This effect suggests that performance on the task of combining forecasts is susceptible to incentives, just like performance on the task of interpreting time-series data, presumably because it requires intensity and duration of effort.

In summary, the results from Experiment 2 provide little evidence that investors will fail to account for redundancy and therefore be too extreme. On the contrary, investors respond quite differently (and appropriately) across the observation conditions and their estimates are, if anything, nominally more extreme in the *no-observation* setting than in the observation setting.

## V. Conclusion

Analysts' ability to observe other analysts' forecasts potentially leads to various problems. In our first experiment, we control for any structural problems that might arise from incentives based on relative performance (which can cause herding) and from limited flexibility in adjusting forecasts (which can cause informational cascades), in order to examine whether observation harms consensus accuracy because of miscalibration or by inducing free-riding. We find little support for concerns about miscalibration. Instead, our results show that mutual observation allows analysts to aggregate their information, which reduces forecast dispersion and increases consensus accuracy and extremity. However, mutual observation does induce free-riding, which reduces the amount of information available to be aggregated. When incentives for accuracy are very high, mutual observation induces little free-riding, so that mutual

observation has a net positive impact on consensus accuracy. When incentives are relatively low, free-riding is more severe and mutual observation harms accuracy.

In our second experiment, we present the analysts' assessments from the first experiment to a group of investor participants. We find strong evidence that investors account for the different statistical properties of consensus forecasts across the two settings. In particular, investors appear to understand that more extreme forecasts should be weighted more heavily in the no-observation setting and not in the observation setting. Contrary to results from some prior experiments, we find that investors dramatically alter how they respond to the consensus, conditional on whether mutual observation was allowed among analysts and do so in a manner consistent with economic theory. Moreover, we find little evidence that allowing mutual observation among analysts caused investors to overreact to the consensus, despite the fact that mutual observation did significantly reduce dispersion and increase dependence among analyst forecasts.

Our evidence on the importance of free-riding differs significantly from the focus of much recent research and regulatory action on analyst forecasts. Prior archival research argues that analysts publish forecasts that differ from their true beliefs because they have incentives to please the companies they cover (e.g., Michaely and Womack 1999) or to generate high trading volume (e.g., Cowen, Groysberg, and Healy 2006).<sup>8</sup> In response, recent regulatory efforts have emphasized separating the analyst and investment-banking departments of firms publishing research reports, and requiring firms with investment-banking relationships to include independent reports along with their

---

<sup>8</sup> Anecdotal evidence supports these accusations. Cassidy (2003) describes the case of Henry Blodget, a former lead analyst at Merrill Lynch. On the same day he initiated coverage on GoTo.com and gave a favorable recommendation, he received an email from a fund manager at American Express asking, "What's so interesting about GoTo except banking fees???" Blodget's response? "Nothin."

own reports on firms they have underwritten (SEC 2003). Such actions, while perhaps helpful in addressing bias, will not remedy the effects of free-riding that we document here. Moreover, we note that, while reporting bias can be undone by investors who are aware of analysts' incentives, investors have no recourse if analyst free-riding reduces the amount of information made publicly available in analysts' reports.

One limitation of our study is that the costs of information collection as well as the benefits for forecast accuracy will likely differ significantly in various real world settings. Consequently, we cannot conclude from our experiments alone whether mutual observation would make the consensus more or less accurate in any given situation. Nonetheless, our results still suggest that variations in those costs and benefits will alter the net effect that mutual observation has on accuracy.

Future research might extend from our research setting by examining how specific features of analyst reports and interactions influence the positive and negative aspects of mutual observation. Analyst reports typically include extensive discussion of the details leading to the forecast and persuasive arguments to support conclusions and recommendations. Analysts' identities (and therefore reputations and social status) are well-known to investors and other analysts. Prior research indicates that these types of factors can harm group performance because they are often tied less closely to performance than the conventional wisdom might suggest. Our results provide a foundation on which future research can incorporate these (or other) features when extending the findings of prior research on forecasting into interactive settings.

## REFERENCES

- Becker, G. M., DeGroot, M. H., and J. Marschak. 1964. Measuring utility by a single-response sequential method. *Behavioral Science* 9, 226-232.
- Bikhchandani, S., Hirshleifer, D., and I. Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *The Journal of Political Economy* 100 (5): 992-1026.
- Bloomfield, R. 2002. The “incomplete revelation hypothesis” and financial reporting. *Accounting Horizons* 16 (3): 233-243.
- Bloomfield, R., Libby, R., and M. W. Nelson. 2000. Underreactions, Overreactions and Moderated Confidence. *Journal of Financial Markets* 3 (2): 113-137.
- Bonner, S. E., and G. B. Sprinkle. 2002. The effects of monetary incentives on effort and task performance: Theories, evidence, and a framework for research. *Accounting, Organizations and Society* 27: 303-345.
- Cassidy, J. 2003. The investigation. *The New Yorker* 79 (7): 54-73.
- Clement, M. B., and S. Y. Tse. 2005. Financial analyst characteristics and herding behavior in forecasting. *Journal of Finance* 60 (1): 307-341.
- Cowen, A., Groyberg, B., and P. M. Healy. 2006. Which type of analyst firms are more optimistic? *Journal of Accounting & Economics* 41 (2006): 119-146.
- Dawes, R. M. 1979. The robust beauty of improper linear models in decision making. *American Psychologist* 34 (7): 571-582.
- Dechow, P. M., Hutton, A. P., and R. G. Sloan. 1999. An empirical assessment of the residual income valuation model. *Journal of Accounting and Economics* 26: 1-34.

- Dreman, D. 2002. Bubbles and the role of analysts' forecasts. *The Journal of Psychology and Financial Markets* 3 (1): 4-14.
- Einhorn, H. J., and R. M Hogarth. 1975. Unit weighting schemes for decision making. *Organizational Behavior and Human Performance* 13: 171-192.
- Francis, J., and D. R. Philbrick. 1993. Analysts' decisions as products of a multi-task environment. *Journal of Accounting Research* 31 (2): 216-30.
- Frankel, R., and C. M. C. Lee. 1998. Accounting valuation, market expectation, and cross-sectional stock returns. *Journal of Accounting and Economics* 25 (June): 283-320.
- Glassman, J. K. 2001. When trust collides with risk. *The Washington Post*, December 9.
- Gigerenzer, G., Hoffrage, U., and H. Kleinbolting. 1991. Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review* (98): 506-528.
- Hong, H., Kubik, J.D., and A. Salomon. 2000. Security analysts' career concerns and herding of earnings forecasts. *RAND Journal of Economics* 31: 121-144.
- Hunton, J. E., and R. A. McEwen. 1997. An assessment of the relation between analysts' earnings forecast accuracy, motivational incentives and cognitive information search strategy. *The Accounting Review* 72 (4): 497-515.
- Hulbert, M. 2002. New rules are giving analysts a herd mentality. *The New York Times*, December 29.
- Isenberg, D. J. 1986. Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology* 50 (6): 1141-1151.
- Janis, I. L. 1982. *Groupthink* (2<sup>nd</sup> ed.). Boston: Houghton-Mifflin.

- Jegadeesh, N., and W. Kim. 2006. Imitation or information-driven herding? An analysis of analysts' recommendations and market reactions. Working paper, Emory University.
- Kargin, V. 2003. Prevention of herding by experts. *Economics Letters*, 78 (3), 401-407.
- Krische, S. D. 2005. Investors' evaluations of strategic prior period benchmark disclosures in earnings announcements. *The Accounting Review* 80 (1): 243-268.
- Larrick, R. P., and J. B. Soll. 2006. Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science* 52 (1): 111-127.
- Lynch, A. 2002. Thought contagions in deflating and inflating phases of the bubble. *The Journal of Psychology and Financial Markets* 3 (2): 112-117.
- Libby, R., and M. G. Lipe. 1992. Incentives, effort, and the cognitive processes involved in accounting-related judgments. *Journal of Accounting Research* 30 (2): 249-273.
- Maines, L. A. 1990. The effect of forecast redundancy on judgments of a consensus forecast's expected accuracy. *Journal of Accounting Research* 28 (Supplemental): 29-47.
- Maines, L. A. 1996. An experimental examination of subjective forecast combination. *International Journal of Forecasting* (June): 223-233.
- Maines, L. A., and J. R. M. Hand. 1996. Individuals' perceptions and misperceptions of time series properties of quarterly earnings. *The Accounting Review* 71 (3): 317-336.

- Maines, L. A., McDaniel, L. S., and M. S. Harris. 1997. Implications of proposed segment reporting standards for financial analysts' investment judgments. *Journal of Accounting Research* 35 (Supplement): 1-24.
- Michaely, R., and K. L. Womack. 1999. Conflict of interest and the credibility of underwriter analyst recommendations. *Review of Financial Studies* 12 (4): 653-686.
- Myers, D. G., and H. Lamm. 1976. The group polarization phenomenon. *Psychological Bulletin*. 83: 602-627.
- Opdyke, J. D., and L. Asinof. 2001. Roundtable: The problem with the herd. *The Wall Street Journal*, January 29.
- Ritholtz, B. 2004. "What does Mr. Market want (and should we even care)?" The Big Picture. July 14. (<http://bigpicture.typepad.com/comments/2004/07>). 26 Oct 2005.
- Scharfstein, D.S., and J. C. Stein. 1990. Herd behavior and investment. *American Economic Review* 80 (3), 465-479.
- Securities and Exchange Commission. 2003. SEC Fact Sheet on Global Analyst Research Settlements. (<http://www.sec.gov/news/speech/factsheet.htm>)
- Sedor, L. M. 2002. An explanation for unintentional optimism in analysts' earnings forecasts. *The Accounting Review* 77 (4): 731-753.
- Seybert, N., and R. Bloomfield. 2006. Contagion of "wishful thinking" biases in laboratory markets. Working paper, Cornell University.
- Shiller, R. 2005. *Irrational Exuberance*. 2nd edition. Princeton, NJ: Princeton University Press.

- Soll, J. B. 1999. Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psychology* 38: 317-346.
- Soll, J. B., and R. P. Larrick. 2006. Strategies for revising judgment: How, and how well, do people use others' opinions? Working paper, Duke University.
- Trueman, B. 1994. Analyst forecasts and herding behavior. *Review of Financial Studies* 7 (1), 97-124.
- Welch, I. 1992. Sequential sales, learning, and cascades. *Journal of Finance* 47 (2), 695-732.
- Zwiebel, J. 1995. Corporate conservatism and relative compensation. *The Journal of Political Economy* 103 (1), 1-25.

**TABLE 1**  
**Firm Information**

Firm	ROE History							Optimal		Quartile Rank	
	2	3	4	5	6	7	8	Alpha	V/B	Alpha	Performance
1.1	11.6	11.1	10.6	9.2	13.1	8	-	41.5%	0.86	1	1
1.2	11.6	11.1	10.6	9.2	13.1	8	14.1	36.5%	1.42	2	2
2.1	17.5	15.7	20.5	19.6	21.1	19.2	-	81.9%	5.62	3	4
2.2	17.5	15.7	20.5	19.6	21.1	19.2	21.2	89.8%	6.05	4	4
3.1	10.9	10.2	11.1	13.2	9.4	3.7	-	48.0%	0.63	1	1
3.2	10.9	10.2	11.1	13.2	9.4	3.7	17.2	17.8%	1.53	4	3
4.1	13.4	20.1	15.6	16.6	16.7	6	-	31.5%	0.80	2	1
4.2	13.4	20.1	15.6	16.6	16.7	6	12.4	23.2%	1.15	3	2
5.1	6.7	11.5	13.6	12.8	16.1	18.8	-	61.2%	3.17	2	3
5.2	6.7	11.5	13.6	12.8	16.1	18.8	22.8	80.3%	7.06	3	4
6.1	13.2	22.1	23.6	26.5	22	20.1	-	93.3%	5.16	4	4
6.2	13.2	22.1	23.6	26.5	22	20.1	15.3	89.2%	2.32	4	3
7.1	5.2	8.7	7.5	7.8	1.3	9	-	29.5%	0.94	2	1
7.2	5.2	8.7	7.5	7.8	1.3	9	15.1	27.3%	1.44	3	2
8.1	11.8	10.4	15.7	16.4	14	12.6	-	50.6%	1.31	1	2
8.2	11.8	10.4	15.7	16.4	14	12.6	16.6	56.0%	2.20	1	3

This table contains information about each of the firms used in the experiment. Each firm was analyzed for two rounds – once with a seven-period history and once with an eight-period history. All histories began with ROE = 10%, the cost of capital. Participants moved through the eight “firms” in numerical order, as indicated above, or in the reverse order. Alpha refers to the probability that abnormal earnings persist at a high rate (95%) rather than the low rate (65%). Optimal alpha is the statistical probability that alpha is high, given the ROE history. Alpha diagnosticity measures how far the optimal alpha is in absolute terms from the naïve prior of 50%. Similarly, optimal value-to-book measures the VB ratio implied by the optimal alpha, and VB diagnosticity measures how far the optimal VB ratio is, in absolute terms, from the VB ratio implied by the naïve alpha of 50%. Performance rank is determined by ranking the 16 events according to the absolute magnitude of residual income and classifying them into quartiles, where quartile four reflects the most extreme quartile of abnormal performance.

**FIGURE 1**  
**Screen Display**

*Panel A: Tabular Display of Financial Information*

<u>Estimates</u>	<u>Mine</u>	<u>Consensus</u>	<u># of Tentatives</u>
Current BV		0.00	0
Current ROE		0.00	0
Probability of HIGH Persistence		0.00	0
Value-to-Book Multiple	0.00		

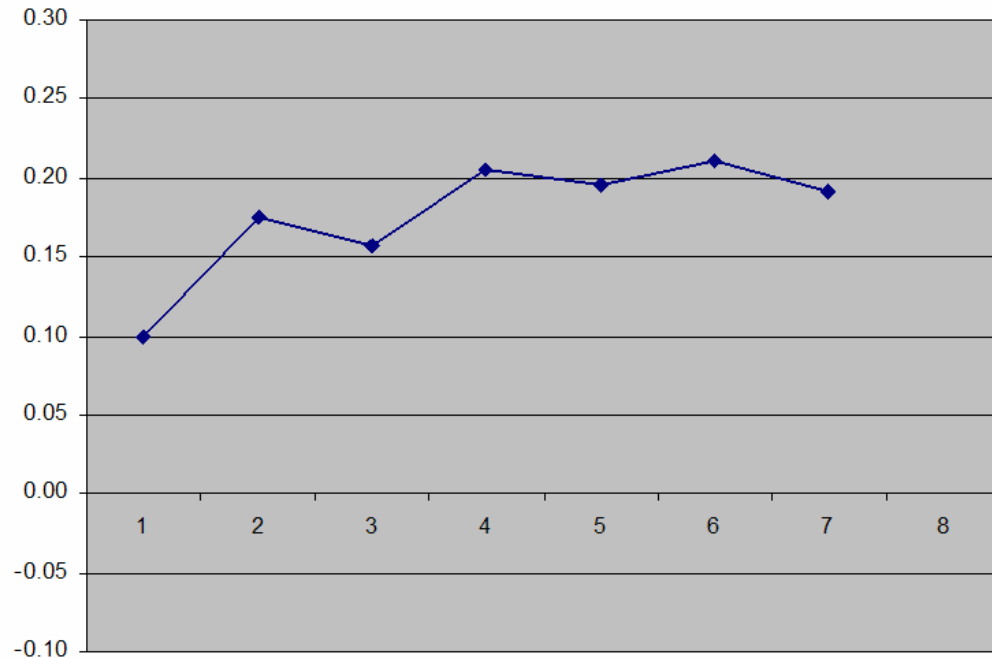
  

Synthetic Value Estimate	<u>Mine</u>	<u>Consensus</u>	Persistence is either...	
			High	Low
	0.00	0.00	0.95	0.65

<u>ANALYST INFO</u>	<u>Beginning of Year</u>	<u>Earnings</u>	<u>ROE</u>
<u>YEAR</u>	<u>BV History</u>	<u>History</u>	<u>History</u>
1	100.00	10.00	10.00
2	110.00	19.22	17.50
3	129.22	20.35	15.70
4	149.57	30.59	20.50
5	180.16	35.25	19.60
6	215.41	45.42	21.10
7	260.84	49.96	19.20
8	310.80	-	-
9	-	-	-

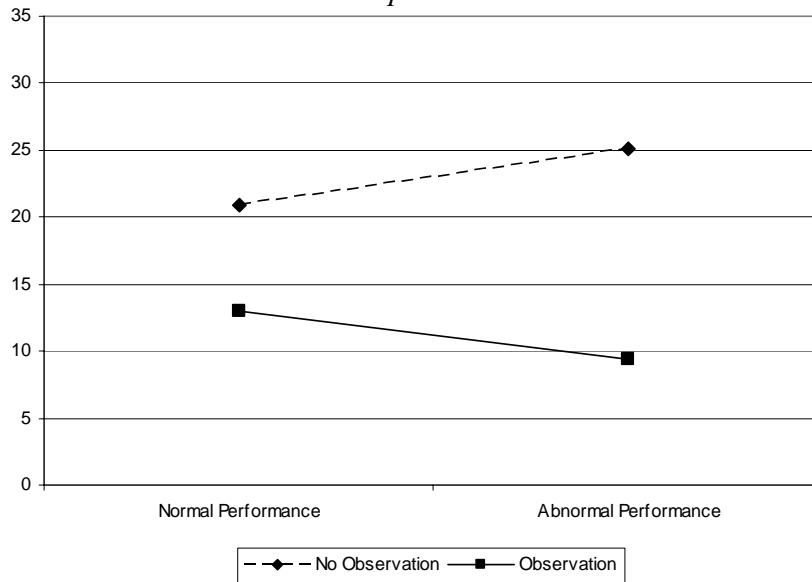
*Panel B: Graphical Display of Historical ROE*



This figure contains snapshots of the information as displayed on the computer screen. ROE history was displayed both in tabular format and graphically.

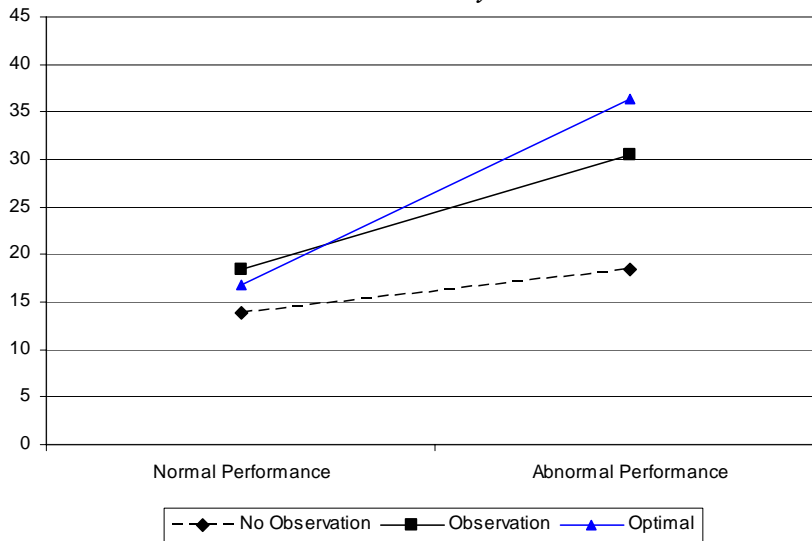
**FIGURE 2**  
**Forecast Properties Conditional on Observation and Abnormal Performance**

*Panel A: Forecast Dispersion*



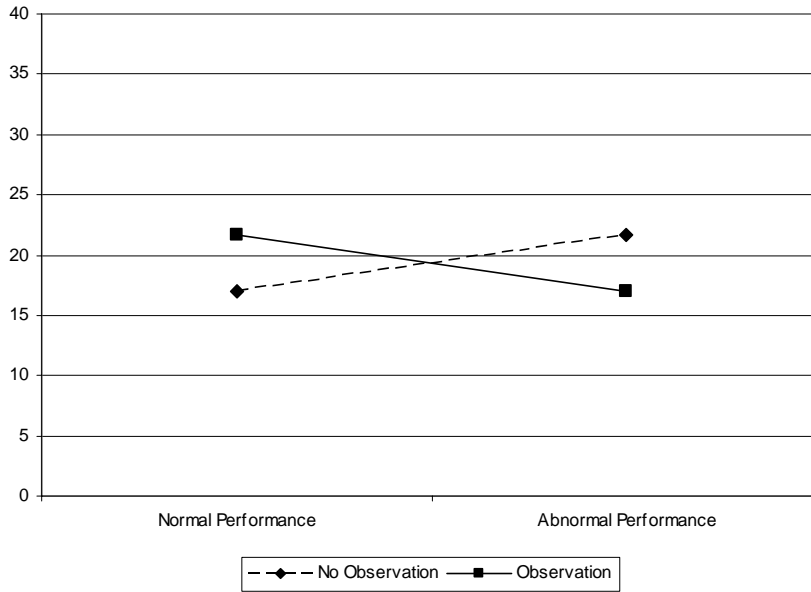
Observation:	$p < 0.001$
Diag:	$p = 0.710$
Performance:	$p = 0.576$
Obs x Diag:	$p = 0.267$
Obs x Perform:	$p < 0.001$

*Panel B: Consensus Extremity*



Observation:	$p = 0.008$
Diag:	$p = 0.702$
Performance:	$p < 0.001$
Obs x Diag:	$p = 0.153$
Obs x Perform:	$p = 0.008$

Panel C: Consensus Error



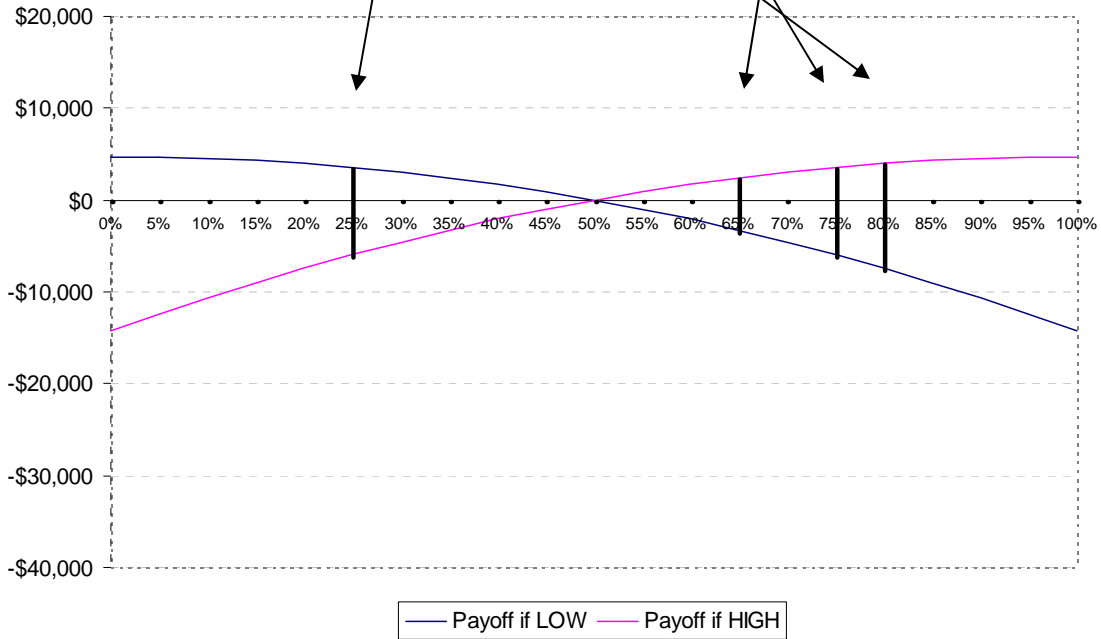
Observation:	$p = 0.838$
Diag:	$p < 0.001$
Performance:	$p < 0.001$
Obs x Diag:	$p = 0.577$
Obs x Perform:	$p = 0.005$

This figure reports consensus dispersion, extremity, and absolute error of participants' alpha estimates, conditional on whether mutual observation is allowed and whether firm performance is abnormal or not. Within-group forecast dispersion is measured as the absolute deviation of each forecast in a cohort from that cohort's mean (consensus) forecast. Consensus extremity measures the absolute deviation between the consensus forecast and the naïve prior of 50%. Consensus error measures the absolute deviation between the consensus forecast and the optimal alpha implied in the data series. Alpha refers to the probability that abnormal earnings persist at a high rate (95%) rather than the low rate (65%). No-observation groups consist of four individuals who submit their assessment of alpha without learning anything about the forecasts of their fellow group members. Observation groups consist of four individuals who have access to the current consensus forecast for their group prior to submitting their final forecast. To the right of each panel, we report  $p$ -values from the full repeated-measures ANOVA described in the text.

**FIGURE 3**  
**Example Table and Graph from Experiment 2**

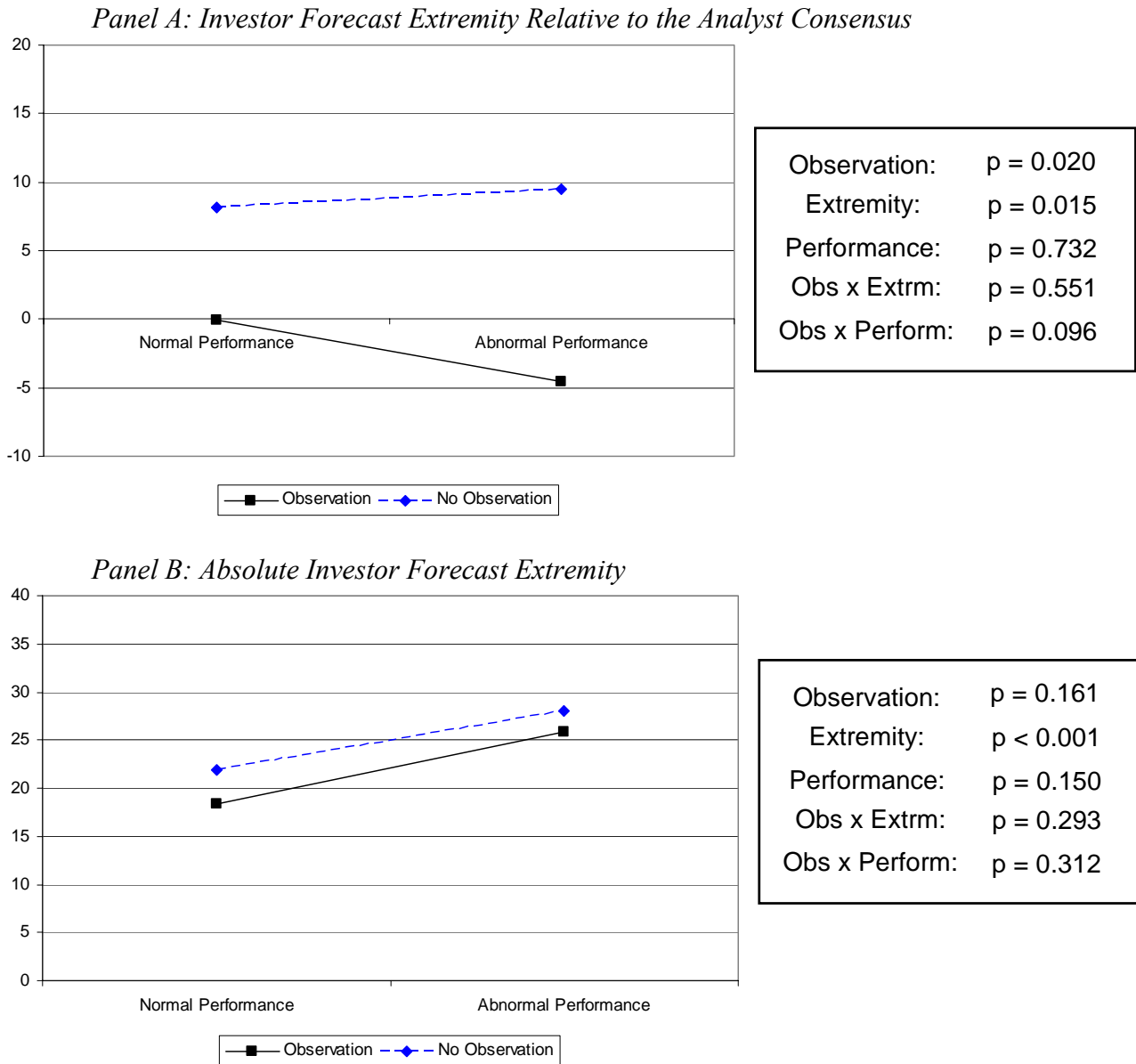
	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>Mean</b>	<b>Median</b>
<b>Judgments</b>	<b>25%</b>	<b>80%</b>	<b>75%</b>	<b>65%</b>	<b>61%</b>	<b>70%</b>
<i>Payoff if Persistence is Low</i>	3,565	-7,416	-5,942	-3,280	-2,322	-4,564
<i>Payoff if Persistence is High</i>	-5,942	3,993	3,565	2,424	1,862	3,042

**Payoff Graph**



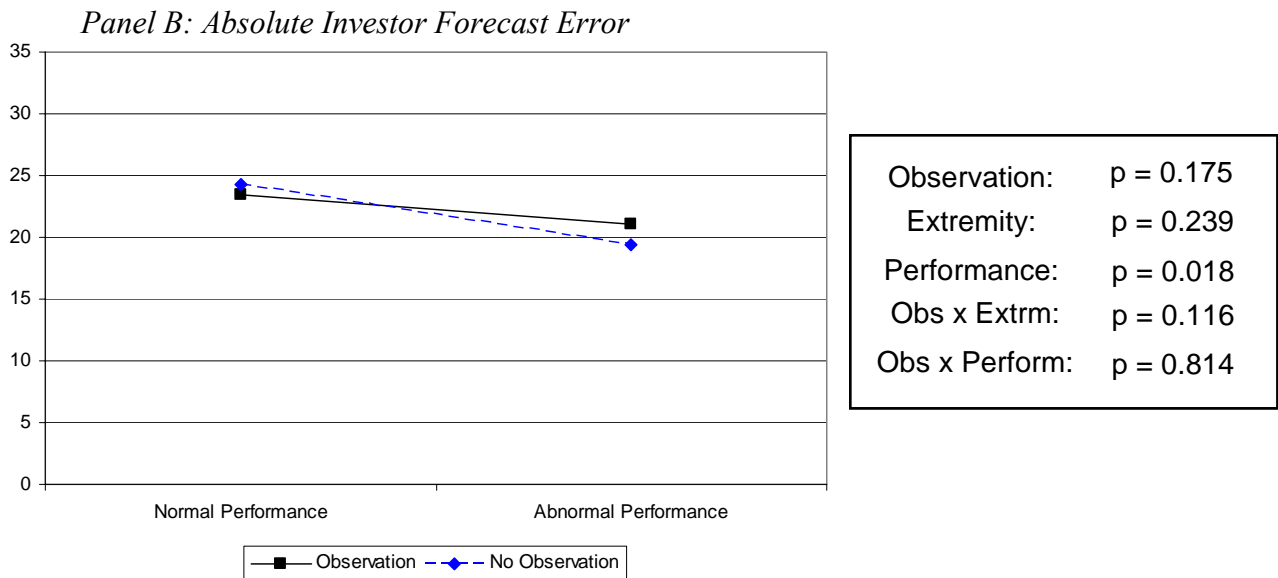
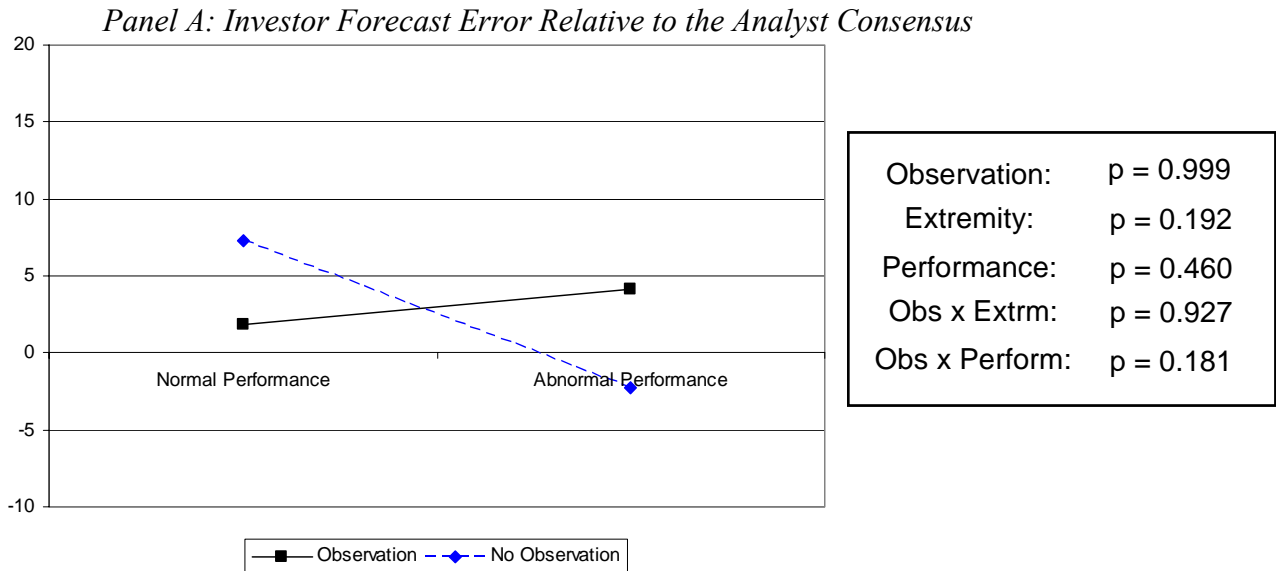
This figure contains an example table and graph from Experiment 2. The table contains assessments of the probability that persistence is HIGH for this firm. The payoff graph translates the individual forecaster assessments into their payoff implications, conditional on persistence actually being high or low.

**FIGURE 4**  
**Investor Forecast Extremity**



This figure reports the relative and absolute extremity of investors' alpha assessments, conditional on whether mutual forecaster observation was allowed and whether firm performance is abnormal or not. Relative forecast extremity measures the difference between the investor's assessment of alpha and the consensus estimate of the cohort to which the investor had been yoked. Absolute forecast extremity measures the difference between the investor's assessment of alpha and 50%. Alpha refers to the probability that abnormal earnings persist at a high rate (95%) rather than the low rate (65%). In the no-observation setting, investors are yoked to analyst cohorts from Experiment 1 in which mutual observation was not allowed. In the observation setting, investors are yoked to analyst cohorts from Experiment 1 in which mutual observation was allowed. To the right of each panel, we report  $p$ -values from the full repeated-measures ANOVA described in the text.

**FIGURE 5**  
**Investor Forecast Error**



This figure reports the relative and absolute error of investors' alpha assessments, conditional on whether mutual forecaster observation was allowed and whether firm performance is abnormal or not. Relative forecast error measures the absolute deviation between the investor's assessment of alpha and the optimal alpha implied by the data series less the same error measures as calculated using the consensus estimate. Absolute forecast error measures the absolute deviation between the investor's assessment of alpha and the optimal alpha implied by the data series. Alpha refers to the probability that abnormal earnings persist at a high rate (95%) rather than the low rate (65%). In the no-observation setting, investors are yoked to analyst cohorts from Experiment 1 in which mutual observation was not allowed. In the observation setting, investors are yoked to analyst cohorts from Experiment 1 in which mutual observation was allowed. To the right of each panel, we report  $p$ -values from the full repeated-measures ANOVA described in the text.