MANAGEMENT SCIENCE
Vol. 56, No. 10, October 2010, pp. 1653–1667
ISSN 0025-1909 | EISSN 1526-5501 | 10 | 5610 | 1653

DOI 10.1287/mnsc.1100.1220

Predicting the Next Big Thing: Success as a Signal of Poor Judgment

Jerker Denrell
Saïd Business School, University of Oxford, Oxford OX1 1HP, United Kingdom, jerker.denrell@sbs.ox.ac.uk

Christina Fang
Department of Management, Stern School of Business, New York University, New York, New York 10012, cfang@stern.nyu.edu

Successfully predicting that something will become a big hit seems impressive. Managers and entrepreneurs who have made successful predictions and have invested money on this basis are promoted, become rich, and may end up on the cover of business magazines. In this paper, we show that an accurate prediction about such an extreme event, e.g., a big hit, may in fact be an indication of poor rather than good forecasting ability. We first demonstrate how this conclusion can be derived from a formal model of forecasting. We then illustrate that the basic result is consistent with data from two lab experiments as well as field data on professional forecasts from the *Wall Street Journal* Survey of Economic Forecasts.

Key words: managerial foresight; forecasting; resource based view of the firm
History: Received April 13, 2008; accepted May 17, 2010, by George Wu, decision analysis. Published online in *Articles in Advance* September 3, 2010.

1. Introduction

Managers and entrepreneurs are often assessed on their ability to forecast the success of new ventures. Managers evaluate new products and ideas, make bets on which of them will succeed (Harrison and March 1984, Ghemawat 1991, Adner and Helfat 2003), and advance in their careers by predicting successful new products and technologies (March and March 1977, Kidder 2000). Entrepreneurs who can "see what is next" and successfully invest in the "next big thing" (Christensen et al. 2004) become rich and may end up on the cover of business magazines.

Successfully predicting that something will become a big hit seems impressive, and the individuals who get it right are often hailed as "seers" (Armstrong 1978). Underlying the admiring accounts of farsighted individuals is the assumption that managers and entrepreneurs who accurately predicted that a new venture would be successful (i.e., the next big thing), are likely to be better forecasters. Intuitively, an accurate forecast is more likely to have been made by a forecaster who has better judgment and is better able to evaluate the situation. Here, however, we argue that there is a simple reason why this intuition may be wrong. Rather than being an indication of good judgment, accurately forecasting a rare event such as business success may in fact be an indication of poor judgment. The reason is that a forecaster with poor judgment is more likely than a forecaster with good judgment to predict the rare and extreme event of a product becoming successful.

To develop our argument, we employ a three-pronged approach: we build an analytical model and test the implications of the model on both experimental and field data. Using a simple model to formalize our intuition, we examine managers making predictions about the value of a new product based on a noisy signal. We assume that they follow different strategies when making their predictions (Makadok and Walker 2000). Some may rely on systematic approaches, whereas others depend on heuristics and intuition (Kahneman and Tversky 1973, Kahneman and Lovallo 1993). We show analytically that if a manager predicted that an event would be extremely successful and the prediction turns out to be correct, this manager may in fact have poor forecasting ability. In other words, an accurate judgment can be a signal of poor judgment.

The explanation is that because extreme outcomes are very rare, managers who take into account all the available information are less likely to make such extreme predictions, whereas those who rely on heuristics and intuition are more likely to make extreme predictions. As such, if the outcome was in fact extreme, an individual who predicts accurately an extreme event is likely to be someone who relies on intuition, rather than someone who takes into account all available information. She is likely to be someone who raves about *any* new idea or product (Denrell 2005). However, such heuristics are unlikely to produce consistent success over a wide range of forecasts. Therefore, accurate predictions of an extreme event are likely to be an indication of poor overall forecasting ability, when judgment or forecasting ability is defined as the average level of forecast accuracy over a wide range of forecasts.

We test the empirical implications of the model using experimental data we gathered from two lab experiments in addition to field data on professional forecasts from the *Wall Street Journal* Survey of Economic Forecasts. Consistent with our model, both the experimental and field results demonstrate that in a data set containing all predictions, an accurate prediction is an indication of good forecasting ability (i.e., high accuracy on all predictions). However, if we only consider extreme outcomes, then an accurate prediction is in fact associated with poor forecasting ability.

Our results suggest that inferring forecasting ability from a selective set of observations, such as cases of business success, may be more complicated than previously believed. Rather than being impressive, accurate predictions about such extreme events may be an indication of poor forecasting ability.

2. Model

2.1. Model Details

To formalize the intuition that an accurate prediction can be an indication of poor judgment, we construct an analytical model by extending a standard model in which a manager has to make a prediction on the basis of a noisy signal (e.g., Marshak and Radner 1972, Harrison and March 1984). We incorporate the possibility that the manager makes use of intuitive heuristics in formulating the prediction (Kahneman and Tversky 1973). We then compare the accuracy of the forecasts made by a Bayesian manager and a manager who deviates from Bayes's rule.

2.1.1. Prediction Task. Managers make predictions about the success of a new product based on information available about the specific case at hand (e.g., product characteristics, the current competitive outlook, etc). In addition, they also possess information about the base rate of success in their business. Consider, for example, managers at television networks who routinely predict the popular appeal of various proposed series and shows. These predictions are often based on pilot test results that are not completely reliable (Bielby and Bielby 1994, Gitlin 2000, Kennedy 2002) and only provide a noisy signal of future demand. In addition, managers are aware that very few shows become hits (Bielby and Bielby 1994). Therefore, to formulate a prediction, managers need to integrate their prior knowledge about the base rate of success (Kahneman and Lovallo 1993) with the information about the case at hand (which may represent a noisy signal of the underlying variable).

One standard model of such a prediction task based on a noisy signal is the following: a manager who observes a noisy signal, $S = \mu + \varepsilon_1$, of the true performance, μ, of a product. We assume that the manager cannot observe μ directly but knows the distribution of μ within her business. Specifically, we assume that she knows that the true performance level μ is normally distributed, with mean zero and variance $\sigma_u^2 = 1$. Thus, although the manager does not know the specific value of μ that characterizes the new product, she does know that this value of μ is drawn from a normal distribution with a mean of 0 and a variance of 1. Furthermore, the signal she observes contains an error term, ε_1, which is also assumed to be normally distributed with a mean of 0 and variance of $\sigma_1^2 = 1$.

Based on the signal, S, she makes a prediction p about the actual performance level A, where $A = \mu + \varepsilon_2$ (ε_2 is an error term, independent of ε_1, which is also normally distributed with a mean of 0 and variance of $\sigma_2^2 = 1$). The prediction is based on two pieces of information: the observed noisy signal and some prior information about how likely it is that μ is high or low. After the prediction, the actual outcome can be observed (A). We assume that the manager tries to come up with a prediction as close to the actual outcome as possible. Specifically, we assume that she aims to minimize the expected squared difference between the actual outcome and the prediction, $E[(A-p)^2]$, known as mean square error (MSE). A good forecaster is one with a low MSE. Later we discuss how our results may change if the forecaster faces different incentives from the ones assumed here.

2.1.2. Prediction Strategies. Suppose there are only two types of managers following different strategies in formulating their predictions. The first manager is rational and follows Bayes's rule. To minimize the MSE, she sets the prediction equal to the expected value of the posterior (DeGroot 1970).[1] That is, the rational manager sets the prediction p equal to the expected value of the posterior, $E[A \mid S]$, using Bayes's rule:

$$p = E[A \mid S] = E[u \mid S] = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_1^2} S + \frac{\sigma_1^2}{\sigma_u^2 + \sigma_1^2} E[u] \qquad (1)$$

(e.g., De Groot 1970). When $\sigma_u^2 = \sigma_1^2 = 1$ and $E[u] = 0$, it follows that $E[A \mid S] = 0.5S$. By setting $p = 0.5S$, the rational manager takes into account both the prior

[1] In fact, for normal distributions, the loss minimizing prediction is the expected value of the posterior for any loss function that is an increasing function of the absolute distance between the prediction and the outcome. Thus, our analysis holds for a large class of symmetric loss functions.

information (e.g., $E[u]=0$) and the observed signal (whether S is high or low).

We assume that the second manager ignores prior information. A large literature on behavioral decision theory documents that most people do not combine prior information and the observed signal in accordance with the Bayes's rule, but instead rely on the representativeness heuristic (Kahneman and Tversky 1973) and ignore the base rate. Suppose that a TV show receives unusually high ratings scores in initial test runs. If executives rely on the representativeness heuristic, they would infer that this TV show will be a hit because these high initial ratings can be taken to be most representative of a show with considerable promise and wide appeal. As a result, they ignore the low base rate of success (Kahneman and Lovallo 1993), and their predictions are insufficiently regressive. Such failure to take prior information (i.e., base rates) into account properly has been illustrated in numerous experiments (Kahneman and Tversky 1973, Griffin and Tversky 1992, Massey and Wu 2005) and field studies. For instance, Cox and Summers (1987) show that retail buyers' sales projections are insufficiently regressive, relative to historical trends.

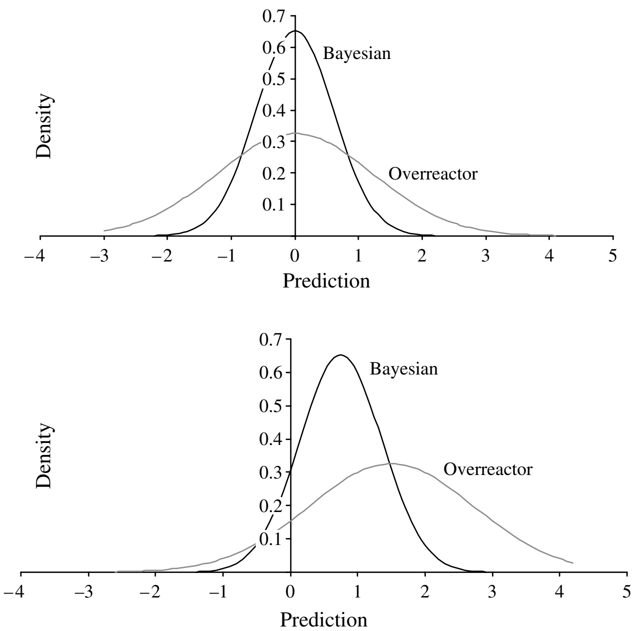
To model such base-rate neglect, we assume that the second manager sets her prediction equal to S; i.e., $p=S$. This prediction strategy leads to a higher expected MSE than the strategy of the first manager, and thus, according to this criterion, the second manager is a worse forecaster. We call the second manager an "overreactor," as she overreacts to the signal by totally ignoring the base rate.

2.2. Results

Figure 1 illustrates how forecasting accuracy varies with the observed outcome. The upper part of the figure plots the distribution of the predictions of the two managers when the actual outcome is 0.[2] Given that μ is drawn from a normal distribution with a mean of 0 and a variance of 1, the actual outcome of 0 is within the reasonable range of expectations. As illustrated, the Bayesian is more likely than the overreactor to make a prediction close to 0. Both distributions are centered on the observed actual outcome of 0. This is because when the actual outcome (A) was 0, the expected value of the signal (S) is $0.5*0$ (the correlation between A and S is 0.5). However, the predictions of the overreactor are more spread out, as she reacts more strongly to any signals. Thus, if the outcome is not extreme, the Bayesian manager is more likely to make an accurate prediction.

The lower part of Figure 1 plots the distribution of the predictions made by the two managers when the actual outcome was 3. Given that μ is drawn from a normal distribution with a mean of 0 and a variance of 1, the actual outcome of 3 represents an extreme outcome. In this case, we see that the Bayesian manager is less likely than the overreactor to make a prediction close to 3. Whereas the predictions of the overreactor are concentrated around $p=1.5$, the predictions of the Bayesian manager are concentrated around $p=0.75$. This is because when the actual outcome (A) was 3, the expected value of the signal (S) was 1.5 (the correlation between A and S is 0.5). The overreactor, who sets the prediction equal to the value of the signal, will make a prediction close to 1.5. The Bayesian, however, makes a prediction that is closer to 0; the expected value of the prediction of the Bayesian is $0.5*E[S]=0.75$. The Bayesian's predictions regress to the mean of the prior distribution (which is 0) because she knows that a high signal is likely to be due to noise, as a high performance level is very rare. Figure 1 therefore illustrates that a forecaster with poor judgment can be more likely to make an accurate prediction when the actual outcome is in fact extreme.

Figure 1 **Distribution of Predictions by Two Managers, Bayesian and Overreactor**

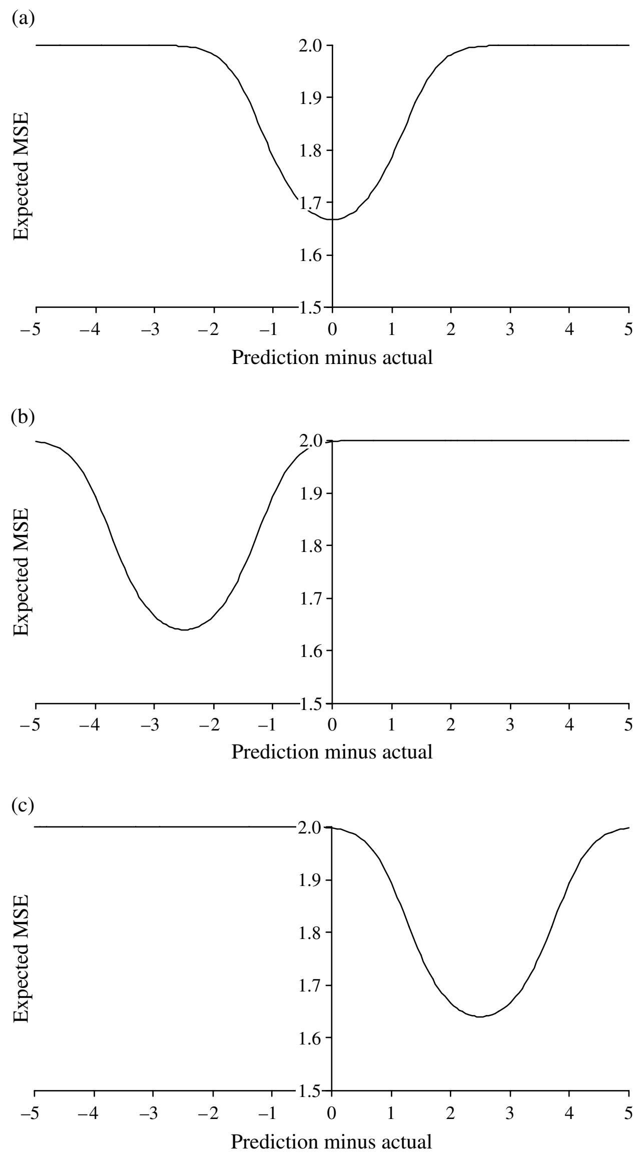


Note. Actual outcome $A=0$ (upper graph) and actual outcome $A=3$ (lower graph).

This implies that in these extreme cases, an accurate prediction of an extreme event is an indication of poor judgment (i.e., a high expected MSE). To illustrate this, we calculated the expected value of MSE for a manager who made a prediction of $P=p$ when the actual outcome was $A=a$; i.e., we calculated $E[\text{MSE} \mid P=p, A=a]$ (see the appendix for details). Figure 2 plots this conditional expectation

[2] Here what is shown is the density of the prediction, given actual outcome $f(P \mid A=a)$. Refer to the appendix for details on how we constructed these graphs.

Figure 2 **Expected MSE as a Function of the Distance Between the Prediction and the Actual Outcome** ($p - A$)



Note. $A = 0$ (panel (a)), $A = 3$ (panel (b)), $A = -3$ (panel (c)).

as a function of the distance between the prediction and the actual value for the different values of the actual outcome ($A = 0$, 3, and -3). One might expect that a prediction equal to the actual outcome would indicate superior forecasting ability (i.e., the lowest expected MSE). However, as shown in Figure 2(a), this is only true if the actual outcome A equals 0. The minimum value of the expected MSE is reached when the prediction is equal to 0. Thus, the most accurate forecasters (i.e., those whose forecast was identical to the actual outcome: $p = A$) are expected to have the highest overall accuracy (i.e., the minimum expected MSE).

A very different pattern is observed when the actual outcome is not equal to zero but is very high. Suppose, for example, that the actual outcome was $A = 3$. In this case, the most accurate forecasters, whose predictions equal the actual outcome, are not expected to be the best forecasters (i.e., with the lowest expected MSE). This is illustrated in Figure 2(b), which shows how the expected MSE varies with the distance between the prediction and the actual outcome when the actual outcome is 3. As shown, the minimum expected MSE occurs when the prediction is lower than the actual outcome ($p < A$). This is because a very high prediction, such as $p = 3$, is an indication of overreaction rather than superior forecasting skills. Suppose, finally, that the actual outcome were very low: $A = -3$. Again, because the actual outcome is extreme, the overreactor is more likely to have made an accurate prediction ($p = A$). Managers who follow Bayes's rule would tend to make predictions that are less extreme and thus closer to 0. As shown in Figure 2(c), the minimum expected MSE occurs when the prediction is higher than the actual outcome ($p > A$).

2.3. Important Boundary Conditions of the Model
With only two types of managers, this simple model shows formally why an accurate prediction about an extreme event might be an indication of poor, rather than good, forecasting skills. There are, however, several important boundary conditions.

First, the assumption that $p = S$ captures an extreme case of base-rate neglect: no weight is given to the prior information (i.e., expected performance level $E[u] = 0$). Usually, people pay some attention to the base rate but weigh it insufficiently relative to the appropriate normative model (Koehler 1996, Hamm 1994, Novemsky and Kronzon 1999). The extent to which information about the base rate is used depends on the context and task. Researchers have demonstrated that direct experience, frequency formats, unambiguous sample spaces, and random sampling tend to promote base-rate usage (Cosmides and Tooby 1996, Gigerenzer and Hoffrage 1995, Tversky and Kahneman 1974, Weber and Borcherding 1993). In reality, managers may vary in the extent to which they take prior information into account. Although our core result holds more generally when there are many managers who vary in their extent of base-rate neglect, it does depend on the frequency of base-rate neglect. Specifically, the basic results would not hold if few or no managers were insufficiently regressive in their predictions.

Consider a model with many different types of managers whose prediction strategies vary in the extent to which they take prior information into account. A manager of type i sets the prediction equal to $p = b_i S$. If $b_i = 0.5$, the manager is a Bayesian. If $b_i > 0.5$, the manager tends to neglect the base rate and overreacts to the signal. If $b_i < 0.5$, the manager pays too much attention to the base rate and underreacts to the signal. Simulations indicate that our basic result, that an accurate prediction about an extreme event signals poor forecasting ability, holds if most managers underutilize the base rate, i.e., have a value of b_i above 0.5. If most managers underreact, however, the opposite result may emerge. Suppose, for example, there were only two managers. The first manager is a Bayesian with a value of b_i equal to 0.5. The second manager is an underreactor with a value of b_i equal to 0.01. The predictions of the second manager will be very close to 0. A high prediction is thus an indication that the individual making the prediction was a Bayesian. As a result, an accurate prediction of a successful activity is an indication of superior forecasting ability (a low expected MSE). However, if we add a third type of manager with a value of b_i equal to 0.9, we get the same result as above. It is interesting to note that in our empirical analysis, presented later, we find that most people tend to underweight the base rate and thus overweight the signal. As a result, most of them have a value of b higher than the Bayesian case of 0.5.

A second critical assumption is that high performance is a rare event, which is reasonable if the model is applied to sales or business success more generally. Because a high outcome is rare, a rational manager is unlikely to predict that the actual outcome will be high even if the observed signal is high. If a high outcome was in fact common, a rational manager would be more likely to predict that the performance would be high if she observed a high signal. For example, suppose that the average outcome (μ) is equally likely to be +3 or −3. Moreover, suppose that the error term is normally distributed with a mean of 0 and a variance of 1. Because the outcome of +3 is a common event, a rational manager will also likely predict a high actual outcome if the signal is high.[3]

Related to this, we assumed that μ and the error term were normally distributed. The results also hold when the distribution of the outcome is (positively) skewed. For instance, most movies sell little, while a few sell a lot (De Vany and Walls 2002). To model this, suppose that sales follow a Poisson distribution with parameter λ. Managers do not know the value of λ but know that it is drawn from a gamma distribution with parameters α and β. Based on test sales (S), managers have to predict the actual sales (A). To do this, a Bayesian would infer the value of λ, based on the test sales, and use this to predict the actual sales (e.g., DeGroot 1970). An overreactor, who ignores the base rate, would only rely on the observed sales and predict that $p = S$. If the distribution of λ is skewed, with most values being small (corresponding to low expected sales for most products), it would be unlikely for a Bayesian to predict high sales even if test sales are high. An overreactor, in contrast, would predict high sales whenever the observed test sales are high. As a result, if the actual sales were high, it is more likely that the overreactor would have predicted such high sales. To illustrate this, we simulated the distribution of predictions, given the actual outcome, for a Bayesian manager and an overreactor. When actual sales $A = 3$ (i.e., close to the most common value of sales when $\alpha = 3$ and $\beta = 1$), we find that the Bayesian manager is more likely to have made a prediction close to the actual outcome. However, when actual sales A are unusually high ($A = 10$), the over-reactor is more likely to have made a prediction close to the actual outcome.[4]

Third, we assumed a particular loss function—that managers wanted to minimize the MSE. Would our basic result hold if forecasters had different incentives, such as one to state a bold and extreme prediction? In short, yes. Even if the rational individual has an incentive to state a prediction higher than the expected value of her posterior, the "irrational" individual would be even more likely to state a very high prediction—because the irrational individual, ignoring the base rate, truly believes that a high outcome is the most likely event given the signal. Stated differently, even if forecasters have incentives to make a bold prediction, the "irrational" forecasters will be more likely to state a bold prediction because they are also more likely to believe that the event will be extreme. Thus, if all forecasters faced incentives to state a bold forecast, we might get the same qualitative results as we have now. The exception is, of course, if incentives to state an extreme prediction are so strong that forecasters ignore the costs of being

[3] In fact, for this alternative model, the rational manager may be more likely than the overreactor to predict a high outcome. If the observed signal is +1, the rational manager would predict an outcome close to +3, whereas an overreactor, who only relies on the signal, would predict +1.

[4] Our results hold more generally when some other assumptions are relaxed. First, the average outcome, μ, does not have to be 0 but can be positive or negative. This does not change the basic results, but only changes the scale. Second, our model assumed that the manager only had access to *one* signal and made only one prediction. Our results would hold but the effect would be attenuated if either condition is relaxed. For instance, if the manager instead has access to more than one signal, she would be able to predict the actual outcome more precisely (this is confirmed in our lab experiments). Similarly, if she makes numerous predictions, the average accuracy of these predictions (the MSE) would be close to the true value (the expected MSE). Thus, an accurate and extreme prediction is most likely to be an indication of poor judgment in situations where only a few noisy signals are available or only a few predictions can be observed.

inaccurate. For example, if all individuals had incentives to always state the highest possible prediction, then their beliefs would not matter for their predictions, and they would behave similarly, regardless of their level of rationality. More generally, strong incentives to state a bold prediction would weaken the association between an accurate extreme prediction and forecast ability, but would not necessarily eliminate or reverse the pattern.

The situation is more complex when forecasters have an incentive to "stand out from the crowd" by stating an accurate and *unique* prediction. Modeling this would require a game theoretic model in which players differ in their level of rationality, and rational players realize this, which is beyond the scope of this paper. Still, it is possible to imagine that such incentives imply that rational forecasters are more likely than irrational forecasters to state an extreme prediction, because rational forecasters realize the need to separate themselves from others.[5] In this case, an accurate but extreme prediction would be an indication of rationality rather than irrationality. Note, however, that this does not change our basic result: an accurate but extreme prediction would still be a signal that the forecaster has high MSE. However, the mechanism is different: rational forecasters realize that they have to sacrifice overall accuracy for the possibility of being unique.

2.4. Related Phenomenon: Regression to the Mean

Our argument is conceptually distinct from "regression to the mean," where extreme values are usually followed by values closer to the mean (e.g., Harrison and March 1984). In our setting, regression to the mean implies that a forecaster who made a very accurate prediction in period t is likely to make a less accurate prediction in period $t+1$. The forecaster is unlikely to repeat an earlier feat, because that might have been due to good luck. This phenomenon occurs in any model in which forecasting accuracy is influenced by chance events, in addition to differences in judgment. However, it is not the mechanism that produces our model's results. In our model, poor forecasters are more likely to make extreme predictions. Therefore, a forecaster who made a very accurate prediction about an extreme event in period t is likely to make a prediction in period $t+1$ that is less accurate than one made by someone who made a less accurate prediction in period t. Hence, we show "regression to *below* the mean": a highly accurate forecaster in one period will have a lower-than-average forecast accuracy in the next period.

[5] This is not obvious, however, because if *many* forecasters have incentives to state a unique prediction and many of them make bold predictions, being unique might require making a conservative prediction.

3. Empirical Illustrations

Our model shows that poor forecasters are more likely to predict accurately when the actual outcome is extreme, because they are more likely to make extreme predictions. To empirically examine the predictions from our model, we conducted two lab experiments. Furthermore, we analyzed a field data set to see whether similar results hold where professional forecasters have substantial expertise and many opportunities to learn and are motivated to be accurate.

3.1. Lab Experiment 1

We asked participants to predict album sales of a series of hypothetical artists based on information from pilot test results. Given the predicted sales and the actual sales, we computed the MSEs for the participants and examined whether an accurate prediction was associated with a high or low MSE.

3.1.1. Participants and Procedures. New York University undergraduates ($N = 133$) were recruited to take part in a computerized experiment, which asked them to play the role of an executive working for a major music label. We told them that their goal was to predict sales of an artist's first album as accurately as possible, for a series of different artists. We also instructed that to do this, they could rely on test results from a "pilot" that evaluated the "sales potential" of an artist. To allow participants to learn about the distribution of the sales, we introduced some hypothetical, historical data on pilot test results and the *actual* first album sales of 25 artists. No further information was provided on each artist, who was identified by a running number only. A randomly drawn sample is shown below:

Artist	Test	Actual
1001	58.62	75.33
1002	27.68	16.17
1003	63.81	70.79
1004	47.86	46.46
1005	43.94	64.67
1006	48.05	44.31
1007	71.04	52.03
1008	54.6	51.37 ...

For instance, artist 1001 attains actual sales of 75.33 (in thousands), whereas the test indicates that her potential is 58.62 (in thousands). Each participant then predicted the actual sales of 100 hypothetical artists. In each case, participants were shown a pilot test result and asked to input their prediction of the actual sales. Afterward, the actual sales would be displayed, as well as the difference between the prediction and the actual sales. The program then showed the test results for the next artist. A window on the screen was created to capture the entire history of the test results, the predictions, and the actual sales for each trial.

For each participant, we randomly generated test results as well as actual sales figures in the following way: for artist i, a random variable, μ_i, was first drawn from a normal distribution with a mean of 50 and a standard deviation of 10. The test sales as well as the actual sales for artist i were then drawn from a normal distribution with mean μ_i and standard deviation of 10. Thus, conditional on μ_i, the test sales and the actual sales were independent random variables. This set-up is identical to the model.

We used two alternative incentive schemes to determine the payoffs of the participants: (1) the "incentive" condition in which subjects are paid in proportion to their MSE and (2) the "fixed-pay" condition in which subjects are only paid for participation ($10) and there are no incentives based on performance. In the "incentive" condition, we told the subjects that their reward would depend on how accurate their predictions were with respect to actual outcomes. The smaller their errors, the larger their rewards. Specifically, participants were told that for each prediction, their accuracy would be measured by the MSE, which is simply the average squared difference between the predicted sales and the actual sales for all the artists. Furthermore, they were told that their payoffs depend on the formulae: Reward = $20 − 0.03* (MSE). Lastly, they were told that they would still earn $3 for participating in the experiment even if their reward ends up negative or zero.[6]

Of the 133 participants who showed up, 69 were randomly assigned to the "fixed-pay" condition and the remaining 64 were assigned to the "incentive" condition.

3.1.2. Results. We calculated the MSE for each participant. In the "incentive" condition, participants' average MSE was 212. This is significantly lower than 276—the average MSE for those in the fixed-pay condition ($t = 5.3880$, $p < 0.001$). With incentives, subjects were more accurate on average.

Recall that, as in Figure 2(a), our model predicts that when the outcome is not extreme, a prediction equal to the actual outcome (i.e., $p_i = a_i$) would indicate superior forecasting ability (i.e., the lowest expected MSE). This means that the minimum value of the expected MSE is reached when the prediction is equal to the actual outcome. However, when the actual outcome is extreme (either high or low), as in Figures 2(b) and 2(c), the most accurate forecasters (for whom the predictions equal the actual outcome) are not expected to be those with the lowest expected MSE. When the actual outcome is extremely high, we predict that the minimum expected MSE should occur when the prediction is lower than the actual outcome ($p < A$), and the opposite is true when the actual outcome is extremely low.

Table 1 **Pooled OLS Regressions Results from Lab Experiment 1**

	(1) All data		(2) $A > 60$		(3) $A < 40$	
Variable	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
			(a) With incentives			
Constant	206.959	<0.001	211.0598	<0.001	206.0792	<0.001
$p_i - a_i$	−0.007	0.851	0.5829	0.007	−0.1554	0.0264
$(p_i - a_i)^2$	0.0230	0.016	0.0324	0.004	0.0229	0.015
N	6,400		1,623		1,371	
R^2	0.023		0.0276		0.0239	
			(b) Without incentives (i.e., fixed pay)			
Constant	265.1659	<0.001	273.8232	<0.001	267.1723	<0.001
$p_i - a_i$	−0.0522	0.541	0.7288	<0.001	−1.1031	0.026
$(p_i - a_i)^2$	0.0388	0.001	0.0417	0.012	0.0637	0.002
N	6,900		1,618		1,779	
R^2	0.0385		0.0262		0.0800	

Notes. All results are based on pooled OLS regressions with MSE as the dependent variable. We pooled across experimental subjects and predictions. Standard errors are clustered on individuals.

Thus, our model suggests that the overall forecasting ability (as measured by MSE) can be explained by two independent variables: (1) the distance between the prediction and the actual outcome, $(p_i - a_i)$ and (2) this distance squared, $(p_i - a_i)^2$. If the actual outcome is not extreme, only the squared term would be significant, indicating that the minimum value of the MSE is reached when the prediction is equal to the actual outcome. However, if the actual outcome is in fact extreme, both the linear and the squared term should be significant. Specifically, when the actual outcome is extremely high, the linear and the squared term should be significantly positive, indicating that the minimum expected MSE occurs for predictions lower than the actual outcome (when $p_i - a_i$ is negative). When the actual outcome is extremely high, the linear term should be significantly negative and the squared term significantly positive, indicating that the minimum expected MSE occurs for predictions higher than the actual outcome (when $p_i - a_i$ is positive).

These predictions are confirmed in our ordinary least squares (OLS) regression analysis with the MSE as the dependent variable and $(p_i - a_i)$ and $(p_i - a_i)^2$ as the independent variables. As seen in column (1) of Table 1, which includes all the data, only the squared term is significant. Thus, in this case, the MSE is at a minimum when the prediction equals the actual outcome ($p_i = a_i$). Column (2) in Table 1 shows the results when we used only the data for which the actual outcome was above 60. Both the squared term and the linear term $(p_i - a_i)$ are significant and positive. Consistent with Figure 2(b), this implies that the minimum expected MSE occurs at a prediction

[6] No participant ended up with a negative score.

that is lower than the actual outcome ($p_i < a_i$). Column (3) in Table 1 documents the result when we used only the data for which the actual outcome was below 40. The squared term is significant and positive and the linear term ($p_i - a_i$) is significant and negative. In accordance with Figure 2(c), this implies that the minimum expected MSE occurs at a prediction that is higher than the actual outcome ($p_i > a_i$).[7] We get similar results if the cutoffs were instead above 55 and below 45 or above and below 50.

To examine whether these results emerge because extreme predictions are associated with base-rate neglect, we estimated the following OLS regression for each participant: $p_{j,i} = a_i + b_i S_{j,i} + \varepsilon_{j,i}$. Here $p_{j,i}$ is the j'th prediction made by participant i and $S_{j,i}$ is the j'th test sale "signal" observed by participant i. The average R^2 for these regressions was 60%, suggesting that this simple model is able to explain a substantial part of the variance in the predictions made. Most participants (82%) had a value of b higher than 0.5 (the value of b that would minimize the expected MSE in this setting), which is consistent with our assumption that most participants put too much weight on the signal. Moreover, the value of b was positively associated with the number of predictions a participant made *above* 60 (the correlation is 0.3248, p-value < 0.01, two-tailed test, $N = 133$) as well as with the number of times a participant *correctly* predicted that the actual sales would be above 60 (the correlation is 0.4615, p-value < 0.001), consistent with our expectation that both the number of extreme predictions and the number of extreme correct predictions are associated with base-rate neglect. Similarly, the estimated value of b was positively associated with the number of predictions below 40 (the correlation is 0.3003) and with the number of times a participant correctly predicted that the actual sales would be below 40 (the correlation is 0.4585).

The number of extreme (above 60 and below 40) and correct extreme predictions was also positively correlated with MSE, but the association was not significant, because the signals and outcomes varied substantially across participants. To eliminate this source of variation, we recruited 47 additional subjects and asked them to predict sales for 50 artists and showed the same 50 pairs of test and actual sales figures to all subjects. In this simpler set up, we found that, as predicted, the MSE was positively correlated with the number of predictions above 60 and below 40 (the correlation was 0.51, p-value < 0.001, two-tailed test, $N = 44$). The reason is that poor forecasters make more extreme predictions: the number of times participants correctly predicted that the actual sales would be *above 60* was positively correlated with the MSE (the correlation is 0.43, p-value < 0.01).[8] A different pattern emerged if we examined the association between the MSE and the number of times participants correctly predicted that the actual sales would be *between* 40 and 60. This association is negative (the correlation is -0.48, p-value < 0.02). Thus, participants who made a larger number of accurate *intermediary* predictions were likely to be forecasters with a lower MSE.

3.2. Lab Experiment 2

To incorporate the possibility that an extreme prediction could be a result of superior information, we designed Experiment 2 so subjects had access to two signals instead of only one. For every prediction we displayed two signals, each randomly drawn from a normal distribution with a mean of 0 and a standard deviation of 10. We again implemented two incentive conditions identical to those in Experiment 1. We recruited 115 additional New York University students, 53 of whom were randomly assigned to the "fixed-pay" condition, and the remaining 62 took part in the "incentive" condition.

We then combined the data from this experiment with the data from Experiment 1 and focused on the "incentive" condition for our analysis.[9] The combined data set includes subjects who vary in their access to information (number of signals) and their prediction strategies (how well they used the information they had access to). Thus, in the data set, an extreme but accurate forecast might not necessarily be an indication of poor forecasting ability. Rather, it may be an indication of access to more information (because people with more precise information should rationally weight the signal more and make more extreme predictions). We hypothesize that in the combined data set our basic result still holds, although the magnitude of the effect should be smaller, because an

[7] Similar results emerge if we do not include the mean squared deviation for prediction i in the computation of the average MSE, when estimating the effect of $p_i - a_i$ and $(p_i - a_i)^2$ on average MSE.

[8] The number of times participants correctly predicted that the actual sales would be *low* was also positively associated with the average MSE. However, this association is only significant (at a p-value of 0.05) if we examine *very low* predictions, such as those below 25. A possible explanation of this lack of symmetry is that participants are more likely to be insufficiently regressive if they observe high test sales than if they observe low test sales. Consistent with this, the average estimated value of b was larger when test sales were above 50 than when test sales were below 50. Ganzach and Krantz (1991) documented a similar asymmetry in regressiveness between high and low values.

[9] As expected, the effect of having two signals was mainly to reduce the average MSE: it was 185.5 for two signals, significantly lower than 211.

Table 2 **Pooled OLS Regressions Results from Lab Experiment 2**

	Model 1, $A > 60$		Model 2, $A > 60$	
Variable	Coefficient	p-value	Coefficient	p-value
Constant	193.3644	< 0.001	231.0293	< 0.001
$p_i - a_i$	0.1208	0.224	0.1817	0.055
$(p_i - a_i)^2$	0.0181	< 0.001	0.0183	< 0.001
Signals			−24.8055	< 0.001
N	3,160		3,160	
R^2	0.0200		0.1237	

Notes. All results are based on pooled OLS regressions with MSE as the dependent variable. We pooled across experimental subjects and predictions. Standard errors are clustered on individuals.

accurate but extreme prediction is a weaker indicator of poor forecasting ability.[10]

To examine this, we redid our regressions with MSE as the dependent variable and $p_i - a_i$ and $(p_i - a_i)^2$ as independent variables, using the combined data set (after removing an outlier participant with a MSE in excess of 400). Previously, when we looked at high actual outcomes ($A > 60$), the coefficients on both $p_i - a_i$ and $(p_i - a_i)^2$ were positive and significant. The implication was that the MSE is lower for predictions that are lower than the actual outcome (when $p_i - a_i$ is negative). As seen in Table 2, which reports the regressions results for the combined data set, the coefficient on $(p_i - a_i)^2$ is still significant and positive. The coefficient on $p_i - a_i$ is positive but not significant ($p = 0.224$). Most important, if we control for the number of signals subjects received, the coefficient on $p_i - a_i$ increases in size and is close to significant ($p = 0.055$). What this illustrates is that when there is heterogeneity in some other variable that influences the accuracy of predictions (i.e., some people have access to more precise information), the effect is weaker and may not be significant.[11] Thus, our results are likely to be less important in contexts where access to information differ significantly.

3.3. *Wall Street Journal* Survey of Economic Forecasts

Every six months, the *Wall Street Journal* asks about 50 economists and analysts to forecast a set of macroeconomic statistics for the next six months (e.g., gross national product (GNP), inflation, unemployment, exchange rates, etc.). The survey is published biannually in the beginning of January and July. The forecast of each participant is published together with the name of the forecaster, which motivates the participants to be accurate.

3.3.1. Data and Measures. Data on forecasts and actual outcomes are available in the *Wall Street Journal* (as well as in the online edition).[12] Using this data, we extracted forecasts and actual values for all forecasters participating in any of the seven surveys from July 2002 to July 2005. Each survey asked participants to forecast eight different economic items: gross domestic product (GDP), the unemployment rate, the consumer price index, the three-month Treasury bill, the 10-year government note, federal funds, the yen, and the euro. The median number of surveys that the 68 forecasters participated in was 5.

To compare the accuracy of different forecasts with very different scales, we measured forecast accuracy by the absolute percentage deviation between the forecast and the actual outcome: $|p_{i,t} - a_t|/a_t$, where $p_{i,t}$ is the forecast made by forecaster i in period t and a_t is the actual outcome. The overall measure of accuracy for forecaster i was the average absolute percentage deviation, where the average was taken over all forecasts of the forecaster i in all surveys i participated in. Denote this measure of overall accuracy $AvgDev_i$. To measure the accuracy of a particular forecast j made by forecaster i in period t, we simply calculated the percentage deviation from the actual value, $Dev_{i,j,t} = (p_{i,j,t} - a_{j,t})/a_{j,t}$, and the measure of the same value squared, $Dev^2_{i,j,t}$.

To test our model, we need to identify some cutoff, which represents an "extreme" outcome, defined relative to what can be expected (i.e., relative to the prior of a rational Bayesian forecaster). In both the model and the experiment we classified an outcome as "high" or "low" using the distance between the actual outcome and the mean of the distribution, because we knew the distribution from which the actual outcomes and the signals were drawn. In this context, such a measure does not necessarily make sense, because the variables that forecasters are asked to forecast may not be stationary and could change predictably. Thus, using historical data to identify an extreme outcome may be problematic if there are trends in the data known to the forecasters. For example, suppose that historically a GDP growth of below 1% has been unusually "low." However, all forecasters may be aware that during the next period, the GDP growth will probably be very low. Using a cutoff

[10] Our basic result would not hold, however, if managers who are insufficiently regressive simultaneously have access to better information and thus potentially could make a more accurate prediction. There seems to be no reason to suspect, however, that overreactors would systematically have access to more information or to more precise information.

[11] We also examined those predictions for which the actual outcomes turned out to be below 40. Although the square term is always significant and positive, the linear term is negative but only marginally significant, regardless of whether we controlled for the number of signals: $p = 0.056$ and $p = 0.079$, respectively (the corresponding coefficients are −0.1899 and −0.1680).

[12] We are grateful to Rick Larrick and Jack Soll from Duke University for making their database on these forecasts available to us.

Table 3 **Pooled OLS Regressions Results from the *Wall Street Journal* Field Data**

	(1) All data		(2) $Dist_{j,t} > 0.2$		(3) $Dist_{j,t} < -0.2$	
Variable	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value
Constant	0.1656	< 0.001	0.1776	< 0.001	0.1888	< 0.001
$Dev_{i,j,t}$	−0.0051	0.627	0.1268	< 0.001	−0.0269	0.058
$Dev^2_{i,j,t}$	0.0354	< 0.001	0.2226	< 0.001	0.0283	< 0.001
N	2,944		264		323	
R^2	0.039		0.279		0.059	

Notes. The results are based on pooled OLS regression with average percentage absolute deviation ($AvgDev_i$) as the dependent variable. We pooled across forecasters, surveys, and different forecast areas (e.g., unemployment, GNP, etc). Standard errors are clustered on individuals.

of 1% would not be appropriate. Instead, we classified an outcome as "high" or "low" by using its distance from the average prediction made by all forecasters. Specifically, we used the percentage deviation between the average prediction and the actual outcome: $Dist_{j,t} = (a_t - \bar{p}_{j,t})/\bar{p}_{j,t}$, where $\bar{p}_{j,t}$ is the average prediction made by all forecasters in survey t about forecast item j. This measure indicates whether an outcome was unusually high or low relative to what was expected by most forecasters. Previous research has also demonstrated that the average prediction is usually quite good (e.g., Larrick and Soll 2006).

3.3.2. Results. We again estimated the same set of OLS regressions with the average accuracy, $AvgDev_i$, as the dependent variable, as we did using the experimental data. We pooled all the data across the different forecasters and the different surveys. The independent variables were the percentage deviation between the prediction and the actual outcome, $Dev_{i,j,t}$, as well as this term squared, $Dev^2_{i,j,t}$. Table 3 shows the result of these regressions.

Column (1) in Table 3, which includes all the data, shows that only $Dev^2_{i,j,t}$ is significant. Thus, in this case, the estimated value of $AvgDev_i$ is at a minimum when the prediction equals the actual outcome. Column (2) in Table 3 shows the results when we only used the data for which the actual outcome was high relative to the average prediction. As a cutoff point, we used an actual outcome that was 20% higher than the average forecast. So, in column (2) we only include the data for which $Dist_{j,t} > 0.2$. Both $Dev^2_{i,j,t}$ and $Dev_{i,j,t}$ are now significant and positive. This implies that the minimum estimated value of $AvgDev_i$ occurs at a prediction lower than the actual outcome. If the actual outcome is high, an accurate prediction (i.e., a prediction equal to the actual outcome) is not an indication of a high average accuracy (i.e., a low value of $AvgDev_i$). Column (3) in Table 3 shows the result when we only use the data for which the actual outcome was 20% lower than the average prediction ($Dist_{j,t} < -0.2$). Now $Dev^2_{i,j,t}$ is significant and positive and $Dev_{i,j,t}$ is negative, albeit only marginally significant. This implies that the minimum estimated value of $AvgDev_i$ occurs at a prediction that is higher than the actual outcome.[13]

The explanation for this pattern is that poor forecasters made more extreme predictions. To illustrate this, we calculated, for each forecaster, the proportion of all forecasts that were extreme, in the sense that they were more than 20% above or below the average prediction. The proportion of extreme forecasts was positively correlated with $AvgDev_i$ (the correlation was 0.65, p-value < 0.001, two-tailed test, $N = 68$). Because poor forecasters were more likely to make extreme forecasts, they are also more likely to make extreme forecasts that turn out to be accurate. To examine this we calculated, for each forecaster, the proportion of forecasts that were more than 20% above (or below) the average prediction when the actual outcome was more than 20% above (or below) the average prediction. The correlation between this proportion and $AvgDev_i$ was 0.53 (p-value < 0.001, two-tailed test, $N = 68$). Thus, an ability to call many extreme events correctly was an indication of poor judgment. In fact, the analyst with the largest number as well as the highest proportion of accurate and extreme forecasts had, by far, the worst forecasting record (the highest $AvgDev_i$).[14]

To summarize, both our model and experiments suggest that poor forecasters make more extreme predictions because they rely too much on the information at hand and weigh the base rate insufficiently. It is not possible to test this explanation using the field data, but this explanation is consistent with anecdotes about how analysts who beat the consensus forecast accomplished this. Consider the story of the highest ranked forecaster in the last period in our data, Sung Won Sohn, CEO of Hamni Financial Group. He achieved his first place by being one of a few who correctly predicted a high inflation rate when the consensus forecast was low. He credited his unusually high but accurate inflation forecast to an intuition he developed after visiting a California jeans producer. The producer could not keep up with demands for its $250 jeans. According to the *Wall Street Journal*, "He figured 'there must be money out there if people are willing to pay that much' for bluejeans" (Gerena-Morales and Hilsenrath 2006). Such methods do not

[13] We get similar results if we use dummy variables to control for different forecast items (e.g., unemployment, GNP, etc.) and if we use a different cutoff, such as +/−0.15 or 0.1. We also get similar results if we do not include, in the computation of the dependent variable (average percentage deviation), the absolute percentage deviation for forecaster i.

[14] The positive correlation remains, however, if we delete this outlier.

always work; in the preceding two surveys, Sung Won Sohn was ranked 43 and 49 out of 55.

An alternative explanation of our results is that a successful prediction generates overconfidence. Overconfident individuals, who overestimate the precision of their private information relative to publicly available information, are likely to make less accurate predictions. For example, Hilary and Menzly (2006) show that analysts who have predicted earnings more accurately than the median analyst in the previous four quarters tend to be simultaneously less accurate and further from the consensus in their subsequent predictions. They attribute this to overconfidence, which is a variable trait triggered by reactions to past results. In contrast, we assume for our model a distribution of different types of forecasters with fixed traits, where some forecasters react too strongly to the signal they get, possibly because they are overconfident and thus ignore the base rate. It is not obvious how a model of variable overconfidence, triggered by past results, could explain the findings in our study. Remember that we found that an accurate prediction was associated with a high overall accuracy, if we used all the data. An accurate prediction was only associated with low overall accuracy when the accurate prediction concerned an outcome that was extreme. For overconfidence to explain our result, it would have to be a selective form of overconfidence that only operates when an analyst makes an accurate forecast that substantially deviates from others. It is possible that such selective overconfidence exists in conjunction with the mechanism in our model.

Past success and failure in forecasting can also influence the motivation of forecasters. For example, forecasters who have made inaccurate predictions because of bad luck or poor judgment may deliberately try to make bold forecasts, to have some chance of being highly ranked (e.g., Chevalier and Ellison 1997, Leone and Wu 2007). Successful forecasters, in contrast, may become cautious to avoid spoiling their existing reputation (Prendergast and Stole 1996). Such changes in motivation triggered by past results could explain our finding that extreme and accurate predictions are associated with high MSE, but only if previously unsuccessful forecasters do not change their behavior after an extreme and accurate forecast in such a way that their overall MSE becomes low.

A simpler alternative explanation is heterogeneity in incentives. Specifically, suppose all analysts are rational and that some analysts—but not all—have an incentive to "stand out from the crowd." That is, some analysts have incentives to state an accurate and *unique* prediction, whereas others have an incentive to make an accurate prediction. As discussed in §2.3, such a model might generate the same basic result: an accurate prediction of an extreme event would be a signal that the decision maker is a poor forecaster (with a high MSE). The underlying mechanism differs from our model, however. Being accurate about an extreme event is not a signal of irrationality (all decision makers are assumed to be rational). Rather, it is a signal that the decision maker has incentives to stand out from the crowd. If such decision makers are likely to make more extreme predictions, they are also likely to have high MSE.

Without data on the prediction strategies of analysts, it is difficult to separate this account from our model. It is not clear, however, whether analysts are in fact motivated to make bold forecasts. Empirical research shows that analysts are more likely to be fired if they have made bold and inaccurate forecasts (Hong et al. 2000), and theoretical work shows that analysts may instead have incentives to stick to the consensus forecast (Scharfstein and Stein 1990, Trueman 1994).

3.3.3. Political Forecasts. Our experimental as well as empirical results are consistent with Tetlock's (2005) analysis of the accuracy of political forecasts. Tetlock (2005) finds that forecasters who rely on conviction and ideology are more likely to make accurate predictions about extreme events, but only because they more frequently make extreme forecasts. For example, such ideologically motivated forecasters successfully predicted the Yugoslavia war (Tetlock 2005, p. 89). Nevertheless, they also predicted many other extreme events that never materialized: War has yet to break out between Hungary and Romania; the divorce between Czechs and Slovaks was as civilized as these things get; and Russia has not yet invaded the Baltics (Tetlock 2005, p. 89). Because these forecasters tend to make extreme forecasts that stray far from base rates (Tetlock 2005, p. 85), they have higher miss rates as well as false alarm rates. Forecasters who do not rely on ideology make less extreme forecasts and are less likely to accurately predict extreme events, even though their overall accuracy scores are higher (Tetlock 2005, p. 91).

4. Implications

4.1. Implications for Inferences About Forecasting Ability

Forecasting ability should ideally be determined based on all predictions, not only a selected subset of extreme predictions. In many contexts, however, data on extreme events may be more accessible or salient. Our results illustrate the hazards of inferring forecasting ability from such selective subsets of predictions.

Although this point may be easily grasped in hindsight, we believe it has not always been taken into account in discussions of forecasting in strategy and management. Consider, for example, the attention paid, in the press and in many textbooks, to successful entrepreneurs who became successful by investing in and predicting new trends. Researchers or consultants who are interested in the determinants of visionary entrepreneurship would be studying a sample of predictions of mainly extreme events. This can cause misleading inferences about forecasting ability, unless the mechanisms we have described are kept in mind. More generally, our paper is a reminder that, in addition to superior information and luck (Barney 1986), base-rate neglect is a characteristic likely to be common among entrepreneurs who discover new sources of competitive advantage.

Consider, next, discussions of the failures of incumbent firms to predict and react to new, "disruptive" technologies (Bower and Christensen 1996). Few emerging technologies or business models are disruptive, and it is not easy to detect the ones that are (Kaplan et al. 2003). Because the base rate is low, rational forecasters will seldom bet that a new technology is disruptive. Irrational forecasters, who ignore the base rate and overreact to signals, are more likely to make such calls. This suggests that the failure to predict what technologies will become disruptive is not necessarily a sign of poor judgment, flawed mental models, or inertia (Tripsas and Gavetti 2000). Rather, it may be an indication of good judgment.

More generally, our model suggests that poor forecasters will be overrepresented among those who were able to "see what is next" (Christensen et al. 2004) and who were hailed as "seers" (Armstrong 1978). Of course, according to our arguments, poor forecasters will also be overrepresented among those who falsely claimed that a new technology would be disruptive. Such cases are often ignored in empirical studies of disruptive technologies, because these studies usually only examine the reaction of firms to technologies that did become disruptive. However, if a study is conditioned on the occurrence of a surprising, disruptive event, we should not be surprised when reasonable managers are surprised that the event occurred. As our results illustrate, in such situations, rational individuals will appear as inert and non-responsive, whereas irrational individuals will appear agile and responsive.

4.2. Implications for Inferences from Performance Data

In many contexts, forecast accuracy is related to performance. Managers and entrepreneurs who make more accurate forecasts will often make more money. Scholars in strategic management have also long emphasized that the origins of competitive advantage lie in the foresight of managers (Barney 1986, Cockburn et al. 2000, Teece et al. 1997). Firms can only obtain a competitive advantage by recognizing the value of resources before the competition does (Barney 1986, Makadok and Walker 2000, Durand 2003, Denrell et al. 2003, Ahuja et al. 2005).

This raises the question of whether our basic result can be applied to inferences about forecasting ability from performance data. Specifically, could high performance be an indication of poor forecasting ability? The answer is negative if all predictions are weighted equally in a performance metric. Recall that when we looked at all the data, an accurate prediction was an indication of good forecasting ability. Our basic result only emerges if attention is focused on extreme events. Similarly, high performance would be an indication of poor forecasting skills only if more attention is paid to predictions about extreme events or when such predictions are disproportionally weighted in performance metrics.

There is, however, an important class of investment decisions in which performance will only be influenced by predictions about extreme events. These are decisions in which it is only economical to invest if demand is predicted to be above some threshold. Consider a manager who contemplates an investment that will only be profitable if the predicted price exceeds a fixed cost of entry. Before investing, the manager observes a signal of demand. Suppose further that she faces 10 such investment opportunities. If the performance metric is the total amount of profits, it will disproportionably be based on predictions about extreme events. Only predictions above the threshold lead to investment, and only investments can generate positive or negative wealth.

In this case, managers who ignore the base rate will be overrepresented in two groups: those who have invested several times and made money and those who have invested several times and lost money. Rational, profit-maximizing managers who follow Bayes's rule will be overrepresented among the group that seldom invests. Overall, this implies that forecasting skill will be a nonmonotonic function of performance (as measured by total wealth). In particular, very high performance will be an indication of base-rate neglect. Because poor forecasting skills lead, on average, to low profits in this situation, very high performance is also an indication of low expected future profits.[15] In a similar way, very high managerial performance would be an indication of poor rather than good judgment, if the task required making investments in new products or markets and the performance metric was the total wealth created.

[15] The nonmonotonic relationship between performance and capability does not hold if performance is defined as the average money made, across all investments. However, in many cases the goal is to accumulate wealth, rather than to maximize the average rate of return. The latter could, after all, be achieved by turning down many lucrative investment opportunities and only investing when the rate is very high.

This argument (which can easily be formalized) suggests that inferences about entrepreneurial ability from entrepreneurial success are perhaps more complicated than usually believed. It is, however, mainly applicable to settings in which high performance requires identifying and investing in valuable products and ideas, as well as settings where such decisions have to be based on noisy signals. If performance depends mainly on capabilities and skills and does not require making forecasts (such as in many sports), if forecasts can be based on precise signals, or if there are large differences in the quality of information that individuals base their forecasts on (e.g., people differ substantially in their expertise and experience), our argument is less relevant. Second, the implication for expected future performance may also be ambiguous if performance relies on skills in addition to forecasting. For example, suppose performance requires accurate forecasting and good leadership, and the two components are uncorrelated. If we observe someone having very high performance, our argument implies that this individual may in fact be a poor forecaster. In contrast, high performance also indicates excellent leadership skills. The expected level of future performance is thus ambiguous.

Note, finally, that although many scholars have argued that business success can be due to luck and chance events in addition to differences in capabilities (Alchian 1950; Mancke 1974; Barney 1986, 1997; Arthur 1989; Levinthal 1991; Denrell 2004), these prior contributions do not challenge the fundamental idea that success is a signal of high capability; they only imply that success is at most a very noisy signal of high capabilities. In contrast, our argument suggests that success could be an indication of low capabilities.

5. Conclusion and Possible Future Research

Successfully predicting that something will become a big hit seems impressive. The little model presented in this paper and the analyses of the experimental and field data are a reminder that such accomplishments are not necessarily a sign of competence. The model shows that there is another reason to doubt whether a forecaster can repeat a successful prediction, in addition to the possibility that it is a fluke. The person making such a successful prediction, if it was about an extreme event, may have systematically worse judgment than others who made less accurate predictions. The model is simple and leaves out many aspects, yet to the extent that its implications are not well understood, it is possible that we may end up awarding "forecasters of the year" awards to a procession of cranks, seek to learn from entrepreneurs with extreme convictions and poor judgment, and promote managers who overconfidently make a series of extreme predictions relying on intuition but neglecting available data on base rates.

At a theoretical level, our results show that interesting insight may come from models that assume heterogeneity in rationality. When forecasters differ in the extent to which they conform to Bayes's rule, accurate but extreme predictions may be an indication of base-rate neglect. More generally, our model illustrates that "success" (i.e., accurate predictions, in our case) may not be an indication of rationality. Similar questions have been explored in behavioral finance, where it has been shown that noise traders rather than "smart" money can achieve the highest return (DeLong et al. 1990). An interesting direction for future research is to examine whether other measures of success, such as successful market entry or promotions, are a signal of rationality in models where the level of rationality varies. For example, successful market entry may be caused by exceptional insight or overconfidence (Camerer and Lovallo 1999). A rigorous analysis of such competitive contexts—as well as of the case when forecasters have incentives to state a unique prediction to get noticed—would require a game theoretic model in which players differ in their level of rationality and rational players recognize this. Models of cognitive "hierarchies" (Camerer et al. 2004) seem promising for such extensions of our overall approach.

Acknowledgments

This work has been generously supported by the Ewing Marion Kauffman Foundation. The authors are grateful to Rick Larrick and Jack Soll from Duke University for sharing their database on the *Wall Street Journal* Survey of Economic Forecasts. The authors have benefited from comments on a prior draft by William Baumol, Daniel Levinthal, Zur Shapira, and Sidney Winter; and from feedback from seminar participants at Carnegie Mellon University, the University of Pennsylvania, New York University, Stanford University, the Stockholm School of Economics, the Judgment and Decision Making Conference (2005), the EGOS Annual Conference (2006), and the Organization Science Winter Conference (2008). The authors also thank three anonymous reviewers, the associate editor, and the departmental editor, George Wu. All errors are the authors' responsibility.

Appendix

To plot Figure 1, we need the conditional distribution of the predictions, given the actual outcome $f(P \mid A = a)$. First, we calculate the conditional distribution of the signal, given the actual outcome. Recall that the value of the prediction, given the signal, is $b_i S$.

Because A and S are normally distributed random variables, with zero expected values, variances $\sigma_s^2 = \sigma_u^2 + \sigma_1^2$ and $\sigma_A^2 = \sigma_u^2 + \sigma_2^2$, covariance $\mathrm{Cov}(S, A) = \mathrm{Cov}(u + \varepsilon_1, u + \varepsilon_2) = \sigma_u^2$, and correlation $\rho = \sigma_u^2 / \sqrt{\sigma_S^2 \sigma_A^2}$, the conditional density of the signal given the actual outcome is (e.g., Gut 1995, p. 129):

$$f(s \mid A=a, b_i) = \frac{1}{\sqrt{2\pi\sigma_s^2}\sqrt{1-\rho^2}} \cdot \mathrm{Exp}\left\{-\frac{1}{2\sigma_s^2(1-\rho^2)}\left(s-\rho\frac{\sigma_s}{\sigma_A}a\right)^2\right\}. \quad (2)$$

The density of the prediction, which is a linear function of the signal, is thus

$$f(p \mid A=a, b_i) = \frac{1}{b_i} f\left(\frac{1}{b_i}s \,\middle|\, A=a, b_i\right). \quad (3)$$

To plot Figure 2, we need to calculate the expected MSE given the predictions.

Suppose there are two managers, 1 and 2, with different values of b: b_1 and b_2. Suppose it is equally likely that an individual is a manager of type 1 and 2. The probability that manager 1 made a prediction p when the actual outcome was $A=a$, $P(i=1 \mid p, a)$, is then

$$\frac{(1/b_1)f((1/b_1)s \mid A=a, b_1)}{(1/b_1)f((1/b_1)s \mid A=a, b_1)+(1/b_2)f((1/b_2)s \mid A=a, b_2)}. \quad (4)$$

The expected MSE, given b_i, is $\mathrm{MSE}_i = E((A-p)^2 \mid b=b_i)$.

To calculate this, note that, in general, $E(X^2) = \mathrm{Var}(X) + E(X)^2$. Thus,

$$E((A-p)^2 \mid b=b_i) = \mathrm{Var}(A-p \mid b=b_i) + E(A-p \mid b=b_i)^2. \quad (5)$$

Because $E(A-p \mid b=b_i)=0$, $E((A-p)^2 \mid b=b_i)$ equals

$$\mathrm{Var}(A-p \mid b=b_i) = \mathrm{Var}(A \mid b=b_i) + \mathrm{Var}(p \mid b=b_i) - 2\mathrm{Cov}(A, p \mid b=b_i), \quad (6)$$

or $\mathrm{MSE}_i = \sigma_u^2 + \sigma_2^2 + b_i^2\sigma_u^2 + b_i^2\sigma_1^2 - 2b_i\sigma_u^2 = (1-b_i)^2\sigma_u^2 + b_i^2\sigma_1^2 + \sigma_2^2$.

It follows that the expected MSE, given a prediction of p and an actual outcome of $A=a$, is

$$E[\mathrm{MSE} \mid p, a] = P(i=1 \mid p, a)\mathrm{MSE}_1 + P(i=2 \mid p, a)\mathrm{MSE}_2. \quad (7)$$

References

Adner, R., C. Helfat. 2003. Corporate effects and dynamic managerial capabilities. *Strategic Management J.* **24**(10) 1011–1025.

Ahuja, G., R. Coff, P. Lee. 2005. Managerial foresight and attempted rent appropriation: Insider trading on knowledge of imminent breakthroughs. *Strategic Management J.* **26**(8) 791–808.

Alchian, A. 1950. Uncertainty, evolution, and economic theory. *J. Political Econom.* **58**(3) 211–221.

Armstrong, J. S. 1978. *Long-Range Forecasting: From Crystal Ball to Computer*. Wiley, New York.

Arthur, B. W. 1989. Competing technologies, increasing returns, and lock-in by historical events. *Econom. J.* **99**(394) 116–131.

Barney, J. B. 1986. Strategic factor markets: Expectations, luck, and business strategy. *Management Sci.* **32**(10) 1231–1241.

Barney, J. B. 1997. On flipping coins and making technology choices: Luck as an explanation of technological foresight and oversight. R. Garud, P. Nayyar, Z. Shapira, eds. *Technological Innovation: Oversights and Foresights*. Cambridge University Press, Cambridge, UK, 13–19.

Bielby, W. T., D. D. Bielby. 1994. All hits are flukes: Institutionalized decision making and the rhetoric of network prime-time program development. *Amer. J. Sociol.* **99**(5) 1287–1313.

Bower, J., C. Christensen. 1996. Customer power, strategic investment, and the failure of leading firms. *Strategic Management J.* **17**(3) 197–218.

Camerer, C., D. Lovallo. 1999. Overconfidence and excess entry: An experimental approach. *Amer. Econom. Rev.* **89**(1) 306–318.

Camerer, C. F., T. H. Ho, J. K. Chong. 2004. A cognitive hierarchy model of games. *Quart. J. Econom.* **119**(3) 861–898.

Chevalier, J., G. Ellison. 1997. Risk taking by mutual funds as a response to incentives. *J. Political Econom.* **105**(6) 1167–1200.

Christensen, C., S. Anthony, E. Roth. 2004. *Seeing What's Next: Using Theories of Innovation to Predict Industry Change.* Harvard Business School Press, Cambridge, MA.

Cockburn, I. M., R. M. Henderson, S. Stern. 2000. Untangling the origins of competitive advantage. *Strategic Management J.* **21** 1123–1145.

Cosmides, L., J. Tooby. 1996. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition* **58**(1) 1–73.

Cox, A. D., J. O. Summers. 1987. Heuristics and biases in the intuitive projection of retail sales. *J. Marketing Res.* **24**(3) 290–297.

DeGroot, M. 1970. *Optimal Statistical Decisions*. McGraw-Hill Company, New York.

DeLong, B., A. Shleifer, L. Summers, R. Waldman. 1990. Noise trader risk in financial markets. *J. Political Econom.* **98**(4) 703–738.

Denrell, J. 2003. Vicarious learning, undersampling of failure, and the myths of management. *Organ. Sci.* **14**(3) 227–243.

Denrell, J. 2004. Random walks and sustained competitive advantage. *Management Sci.* **50**(7) 922–934.

Denrell, J. 2005. Should we be impressed with high performance? *J. Management Inquiry* **14**(3) 292–298.

Denrell, J., C. Fang, S. Winter. 2003. The economics of strategic opportunity. *Strategic Management J.* **24**(10) 977–990.

De Vany, A., W. D. Walls. 2002. Does Hollywood make too many R-rated movies? Risk, stochastic dominance, and the illusion of expectation. *J. Bus.* **75**(3) 425–451.

Durand, P. 2003. Predicting a firm's forecasting ability: The roles of organizational illusion of control and organizational attention. *Strategic Management J.* **24**(9) 821–838.

Ganzach, Y., D. H. Krantz. 1991. The psychology of moderate prediction: II. Leniency and uncertainty. *Organ. Behav. Human Decision Processes* **48**(2) 169–192.

Gerena-Morales, R., J. E. Hilsenrath. 2006. Pricey jeans give Hanmi CEO leg up to top rank of forecasters. *Wall Street Journal* (January 3) A2.

Ghemawat, P. 1991. *Commitment: The Dynamic of Strategy*. Free Press, New York.

Gigerenzer, G., U. Hoffrage. 1995. How to improve Bayesian reasoning without instruction: Frequency formats. *Psych. Rev.* **102**(4) 684–704.

Gitlin, T. 2000. *Inside Prime Time*, 2nd ed. University of California Press, Berkeley.

Griffin, D., A. Tversky. 1992. The weighing of evidence and the determinants of confidence. *Cognitive Psych.* **24**(3) 411–435.

Gut, A. 1995. *An Intermediate Course in Probability*. Springer-Verlag, New York.

Hamm, R. 1994. Underweighting of base-rate information reflects important difficulties people have with probabilistic inference. *Psycoloquy* **5**(3), http://www.cogsci.ecs.soton.ac.uk/cgi/psyc/psummary?5.03.

Harrison, J. R., J. G. March. 1984. Decision making and postdecision surprises. *Admin. Sci. Quart.* **29**(1) 26–42.

Hilary, G., L. Menzly. 2006. Does past success lead analysts to become overconfident? *Management Sci.* **52**(4) 489–500.

Hong, H., J. Kubik, A. Solomon. 2000. Security analysts' career concerns and herding of earnings forecasts. *RAND J. Econom.* **31**(1) 121–144.

Kahneman, D., D. Lovallo. 1993. Timid choices and bold forecasts: A cognitive perspective on risk and risk taking. *Management Sci.* **39**(1) 17–31.

Kahneman, D., A. Tversky. 1973. On the psychology of prediction. *Psych. Rev.* **80**(4) 237–25l.

Kaplan, S., F. Murray, R. Henderson. 2003. Discontinuities and senior management: Assessing the role of recognition in pharmaceutical firm response to biotechnology. *Indust. Corporate Change* **12**(4) 203–233.

Kennedy, R. E. 2002. Strategy fads and competitive convergence: An empirical test for herd behavior in prime-time television programming. *J. Indust. Econom.* **50**(1) 57–84.

Kidder, T. 2000. *The Soul of a New Machine*. Back Bay Books, Boston.

Koehler, J. 1996. The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behav. Brain Sci.* **19**(1) 1–53.

Larrick, R. P., J. B. Soll. 2006. Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Sci.* **52**(1) 111–127.

Leone, A., J. S. Wu. 2007. What does it take to become a superstar? Evidence from institutional investor rankings of financial analysts. Simon School of Business Working Paper FR 02-12, University of Rochester, Rochester, NY.

Levinthal, D. 1991. Random walks and organizational mortality. *Admin. Sci. Quart.* **36**(3) 397–420.

Makadok, R., G. Walker. 2000. Identifying a distinctive competence: Forecasting ability in the mutual fund industry. *Strategic Management J.* **21**(8) 853–864.

Mancke, R. B. 1974. Causes of inter-firm profitability differences: A new interpretation of the evidence. *Quart. J. Econom.* **88**(2) 181–193.

March, J. C., J. G. March. 1977. Almost random careers: The Wisconsin school superintendency, 1940–1972. *Admin. Sci. Quart.* **22**(3) 377–409.

Marshak, T., R. Radner. 1972. *The Economic Theory of Teams*. Yale University Press, New Haven, CT.

Massey, C., G. Wu. 2005. Detecting regime shifts: The causes of under- and overreaction. *Management Sci.* **51**(6) 932–947.

Novemsky, N. N., S. Kronzon. 1999. How are base-rates used, when they are used: A comparison of additive and Bayesian models of base-rate use. *J. Behav. Decision Making* **12**(1) 55–67.

Prendergast, C., L. Stole. 1996. Impetuous youngsters and jaded old-timers: Acquiring a reputation for learning. *J. Political Econom.* **104**(6) 1105–1134.

Scharfstein, D., J. Stein. 1990. Herd behavior and investment. *Amer. Econom. Rev.* **80**(3) 465–479.

Teece, D., G. Pisano, A. Shuen. 1997. Dynamic capabilities and strategic management. *Strategic Management J.* **18**(7) 509–533.

Tetlock, P. 2005. *Expert Political Judgment*: *How Good Is It*? *How Can We Know*? Princeton University Press, Princeton, NJ.

Tripsas, M., G. Gavetti. 2000. Capabilities, cognition and inertia: Evidence from digital imaging. *Strategic Management J.* **21** 1147–1161.

Trueman, B. 1994. Analyst forecasts and herding behavior. *Rev. Financial Stud.* **7**(1) 97–124.

Tversky, A., D. Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* **185**(4157) 1124–1131.

Weber, M., K. Borcherding. 1993. Behavioral influences on weight judgments in multiattribute decision making. *Eur. J. Oper. Res.* **67**(1) 1–12.