Inside Intel: Sales Forecasting using an Information Aggregation Mechanism*

Benjamin J. Gillen
Caltech

Charles R. Plott
Caltech

Matthew Shum
Caltech

May 29, 2013

Abstract

We evaluate the performance of an information aggregation mechanism (IAM) implemented inside Intel to forecast unit sales for the company. Developed, refined and tested in the laboratory using experimental methods, the IAM is designed specifically to aggregate information. Its implementation at Intel allows us to test its performance in a much more complex field environment. The IAM provides not only a point forecast of future sales but also yields the full distribution of participants' beliefs regarding this variable. This predictive distribution very closely matches the distribution over outcomes at short horizons while slightly underweighting low-probability realizations of unit sales at long horizons. Compared to Intel's "official forecast," the IAM forecasts perform well overall, even though they predate the official forecasts. The forecasts improvements are most prominent at short forecast horizons and in direct distribution channels, where the effective aggregation of individually-held information drives the IAM to be more accurate than the official forecast over 75% of the time.

*The financial support of the Lee Center for Advanced Networking, the Gordon and Betty Moore Foundation, and the Laboratory of Experimental Economics and Political Science is gratefully acknowledged. We thank Dan Zhou for providing excellent research assistance, Erik Snowberg, Allan Timmermann, Michael Waldman, and audiences at Arizona State, MIT, Rice, and IIOC 2013 for helpful discussions and comments.

1 Introduction

In most companies, internal forecasts of key financial and operational indicators provide a crucial performance metric and input into strategic planning decisions and managing market expectations. Typically, these forecasts are derived from the analysis of in-house experts, collecting dispersed information from disparate sources in a process consisting of as much art as science. In this paper, we study the use of a completely different type of procedure – an information aggregation mechanism based on decentralized competition, motivated by economic theory, and refined and tested through experimental economics. The mechanism has been shown to work well in the simple and special cases of laboratory settings. The challenge is to test the robustness of the same mechanism when operating in the much more complex environment of a Fortune 500 company. Will it work at all? Will it be useful for improving the internal forecasting process used in the company?

The goal of an information aggregation mechanism is to collect and aggregate the information held in the form of the subjective intuitions from a disperse collection of people. This task requires developing instruments to quantize information while setting proper incentives that balance the reward of revealing information with the hazards of free riding on others information and thus might lead to successful information aggregation. In a break from more traditional, theory-derived, approaches to mechanism design, the IAM was formulated and refined in a series of complex laboratory experiments. The study of information aggregation in experimental economics laboratories has a long history, providing a valuable base on which to build. The ability of markets to perform the information collection and aggregation functions and also the sensitivity of such performance to the details of the market institution were first observed experimentally by Plott and Sunder (1982, 1988). Similarly, the possibility that markets might be designed to perform the aggregation function and implemented inside a business is well known (Chen and Plott (2002); Plott (2000)).

The mechanism studied in this paper shares some institutional features with some betting processes as well as markets. Because some features are taken from parimutuel betting systems, we call it a *Parimutuel Information Aggregation Mechanism (IAM)*. Major features of this mechanism were developed as a response to shortcomings revealed in laboratory experiments of information aggregation. At the same time, the design of the mechanism also reflects an attempt to avoid features that might inhibit the application of information aggregation mechanisms inside a business environment. Such features are not relevant in most betting systems, which have a goal of entertaining participants rather than aggregating their private information.

We report results from a field experiment in which the IAM is implemented to forecast unit sales activity by Intel. As an international market leader in the hi-tech sector with annual revenues over $50 billion, Intel has one of the most recognizable brand names among American companies and its products are found in virtually all households in the country. Accurate forecasts of product sales are incredibly important both operationally, ensuring sufficient inventory is available for distribution, and financially, managing market expectations for shareholder value. With myriad distribution channels, forecasting product sales for the organization is an incredible task requiring analysts to aggregate information from sales reports, partner forecasts, and management guidance. As such, the requisite information for forecasting is dispersed through the firm among a variety of stakeholders. Adapting a parimutual IAM to this environment provides a more "scientific" approach to consolidating this information.

At Intel, we set up a collection of parimutuel mechanisms to characterize uncertainty in future realizations of units sold for key products. The range of values that possible sale quantities can take is partitioned into a set of non-overlapping intervals, or "buckets." The analysts participating in each mechanism are asked to purchase "tickets" that pay off when the variable of interest takes a value within a given bucket. Analysts are allowed to buy as many tickets as they wish (up to a budget limit described below) and place them freely in any of the buckets. In this way, the distribution of tickets placed across the different buckets yields a natural measure of analysts' beliefs regarding the future realization of the variable of interest. The information aggregation mechanism automatically aggregates these beliefs across analysts, allowing decision makers to easily form "consensus" forecasts while also obtaining a glimpse into the uncertainty underlying these forecasts. In addition to the IAM's aggregated forecasts, we also have access to an internally-prepared "official" forecast that can serve as a benchmark against which to evaluate the mechanism's performance. Comparing the relative effectiveness of these mechanisms provides a novel validation of the IAM in an incredibly complex field setting.

Our main findings are summarized as follows:

- The degree to which the beliefs recovered from the IAM are consistent with rational expectations depends on the amount of noise and information in the forecasting environment.
  - The IAM beliefs over unit sales matches the distribution of realized sales at short forecast horizons where individual information is relatively rich.

  - At longer horizons, the mechanism's forecast distribution tends to understate the dispersion of uncertainty in sales, underweighting low probability events. Borrowing terminology from the literature on parimutuel markets, we find a "reverse favorite-longshot bias" in the beliefs derived from the mechanism.

- Using tests based on rational forecasting, we show that both the official and IAM point forecasts appear reasonably free of systematic distortions, though the last forecast revision before sales are announced is systematically understated.

- In statistical comparisons, the mechanism's expected outcome outperforms the official sales forecast at short horizons and in direct distribution product channels.

  - The ex-post optimal combination of the two forecasts heavily weights the IAM at short horizons and the official forecast at long horizons.
  - The IAM is particularly helpful at controlling loss when decision makers have an asymmetric loss from forecast errors. For example, if realized unit sales falling short of expectations has a greater impact on loss than exceeding expectations has on gains.

The appealing performance of the IAM is apparent not only in its empirical properties, but also in the degree to which Intel has expanded its utilization of the IAM in its forecasting and planning process. Starting from an initial pilot of the mechanism, Intel has continued to expand its implementation of the mechanism to several markets that target an important piece of the business. Further, the organization has explored using IAMs beyond forecasting sales itself, such as mechanisms suited to evaluating new ventures in research and development, project management risks, and a variety of other business problems that rely on information dispersed among a number of stakeholders.

2 Information Aggregation Mechanisms: Building on Experimental Success

The purpose of an IAM is to quantify and collect information that might be held, in the form of vague and subjective intuitions, by dispersed individuals. The hope is that the collection and aggregation of this information produces a combined signal that has more information content than any single signal. A connection between markets and information transmission dates back to the foundations of economics (see Allen and Jordan (1998) for a review of this early development). These theoretical results suggested that markets are capable of collecting and aggregating information, though exactly how that might happen was an open question.

Motivated by this theoretical suggestion, Plott and Sunder (1982, 1988) looked to experiments as tools for examining the possibility. Plott and Sunder (1982) first demonstrated the ability of continuous double auction markets to transfer information from "insiders" who have information about the state to non-insiders who do not. Plott and Sunder (1988) builds on this initial finding, demonstrating further that the information transmission and collection can go beyond the simple transfer of information to a process of aggregating the information contained in multiple, independent sources. That is, market-based systems could effectively transfer "soft" information that exists in the form of intuitions into a quantitative signal consistent with Bayes Law. Of significance to the current design of an IAM, they demonstrated that the ability of markets to perform this task is dependent on the trading instruments available. In particular, markets perform the collection and aggregation well if populated by a complete set of Arrow-Debreu securities.[1].

The first application of a market based IAM inside a business was conducted by Chen and Plott (2002) inside Hewlett Packard Corporation. They implemented a complete set of Arrow-Debreu securities to aggregate information about future sales. The possible sales were divided into states, each state supporting an Arrow-Debreu security, and a continuous double auction market was opened for each of the securities. Since the payoff of the winning security was one and the payoff of losing securities was zero the prices of the complete set of securities could be interpreted as a probability distribution over the states. The mechanism was reported as successful but its use was limited due to the cost of managing and difficulties related to coordinating the markets. However, Plott (2000) developed ideas about how an IAM might be designed to perform information collection and aggregation functions inside organizations.

The design of the IAM reported and studied here is similar to parimutuel betting processes - hence the name parimutuel IAM. In a parimutuel betting system participants buy tickets on states of nature, such as the winner of a horse race, and tickets are sold at a fixed price. The revenue from all ticket sales are accumulated, called the purse, and paid

[1]Information aggregation does not necessarily happen if the market has a single compound security and all agents do not have the same preferences. However the prices in a single compound security are related to the competitive equilibrium based on private information. This property, which is common to a private information equilibrium, was demonstrated experimentally by Plott and Sunder, characterized theoretically by Manski (2006), and appears in complex field applications in the Iowa Electronic Markets reviewed by Berg et al. (2008)

to the holders of tickets on the winning bucket. The odds computed from this process reflect the number of tickets sold for a bucket divided into the size of the purse. There is a strong tendency for the odds to be related to the frequency with which the winner occurs. That tendency, which suggested a principle for a new type of IAM, was clearly established experimentally by Plott et al. (2003).

However, the parimutuel IAM implemented at Intel differs from partimutuel betting systems in important ways. First, tickets are not sold at a fixed price, but rather prices evolves with an exogenous trend, in order to encourage a timely completion of the process. This timing setup is informed by the experiments in Axelrod et al. (2009) that demonstrate the importance of structuring the process to encourage participants to buy their tickets early rather than waiting until the last second in an attempt to free ride on information supplied by others.[2] Second, for purchasing the tickets, participants are allocated a fixed budget of a synthetic currency that had no value other than to buy tickets in the designated IAM. The use of a synthetic currency follows Plott and Roust (2009), and works to mitigate the negative impact of risk aversion on information aggregation.[3] Finally, the mechanism is not self-financing, with management providing a fixed cash prize distributed in proportion to the number of tickets in the winning bucket.

The IAM which we ran in Intel also differs in important ways from prediction markets, which have flourished in recent years.[4] First, and foremost, the tickets placed by IAM participants are not securities, and cannot be traded. Price speculation, which takes place in markets, cannot take place here. This payoff structure differs from prediction markets,

[2] Plott et al. (2003) demonstrated that the tendency to wait until the last second to buy tickets inhibited contributed to the creation of bubbles and retarded successful information aggregation. This property was replicate by Kalovcova and Ortmann (2009). The importance of observing others in a parimutuel betting system context is examined by Koessler et al. (2012). In a very simple parimutuel betting system information is transferred through a process of observing betting. In addition elicitation of formed beliefs through a process of placing bets results in greater accuracy in the sense of a reduction in a long shot bias than does a process of reporting beliefs. Koessler et al. (2008) find that level-k models are helpful in identifying the process.

[3] Risk aversion has a tendency to inhibit participation even though an agent is informed and thus prevents information from getting into the system. Plott and Roust (2009) demonstrate that poor performance of the mechanism is closely related to poor information and to the extent that risk aversion diminishes the quality of information the removal of risk aversion is important.

[4] The Iowa Electronic Markets constituted the first "prediction markets" in the sense that the price of a binary security can be viewed as a probability and used to predict elections. The success of the IEM in predicting elections around the world is very convincing and set the stage for a growing field called prediction markets. Prediction market applications (both existing and potential) are surveyed in Wolfers and Zitzewitz (2006) and Arrow et al. (2008). Internal prediction markets have recently been deployed inside large corporations, including Google (Cowgill et al. (2009)), to gauge employees sentiments on everything from companys performance to general industry issues.

where securities are traded by market participants over time, generating an endogenous price path which conveys the information privately held by the participants.[5] Second, in continuous markets there is no institution to coordinate the timing of trades and, thus, to maximize the information flow among participants. Thin trading in the market can hinder aggregation.[6] The timing features of our IAM (described above) were adopted to mitigate these problems. As such, our IAM is more of a non-trading *communication protocol*, rather than an asset market. Indeed, the timing of the IAM is coordinated to be compatible with the busy schedules of participants. There is a fixed, pre-announced start and end time so that people know when to log in to actively participate. The sessions themselves are timed to hit key points of the Intel Business cycle. By design, the output of the IAM is freshly available to other business processes that use it.

Third, prediction markets typically rely on self-selection of participants into the market, guided by the principle that increasing the size of the crowd maximizes its wisdom though suffering the costs of a biased sample of participants.[7] In the Intel IAM, however, management invited participants chosen for the information to which they had access given their position in the organization. Finally, we set up IAM's for sales forecasting which can elicit participants' beliefs about variables (unit sales) which can take many (>>2) values; specifically, we set up a complete set of simultaneous instruments, one for each value that the variable can take. This contrasts with many prediction markets, in which the outcome of interest is binary (or otherwise takes a small number of values); for instance, whether Obama or Romney would win the latest presidential election. Our approach operationalizes a general principle (see Plott (2000)) that the extent of information aggregation is limited by the dimensionality of the "message space" in which market participants operate. Taken as a whole, the activity in all these markets yields a complete probability distribution over the event space that, ideally, will reflect the aggregation of private information about the various possible outcomes.

[5]See Manski (2006) for a discussion of the difficulties of interpreting prediction market prices when participants may have heterogeneous beliefs. Moreover, with heterogeneous beliefs, trade may be curtailed or even eliminated for information reasons (Milgrom and Stokey (1982)). Since our IAM is not a trading market, we avoid these possibilities.

[6]This was an issue encountered in the Hewlett-Packard IAM implemented by Chen and Plott (2002).

[7]This phenomenon may be accentuated in horse-racing parimutuel markets, in which individual decisions may be directed by the thrill of uncertainty and surprise rather than the desire to profit from exclusive information. While Woodland and Woodland (1994) and Gray and Gray (1997) find that thick betting markets for professional sports tend to satisfy market efficiency, a host of papers have explored potential cases of inefficiencies in recreational betting markets. Jullien and Salanie (2000) and Chiappori et al. (2009) discuss the identification and estimation of risk preferences using data from parimutuel markets.

3 The IAM inside Intel

For description purposes we will consider a single variable, say unit sales for product i in quarter t, that we denote by $Y_{i,t}$. The positive real line is partitioned into K intervals, or "buckets," where each interval represents a range of possible values for sales that will be officially reported at the end of the sales period. The leftmost and rightmost buckets are, respectively, $[0, x_1)$ and $[x_{K-1}, \infty)$.

Participants interact with the mechanism in the form of an on-line interactive program. Mechanism organizers invite participants, who securely log in to their own account to access the IAM program. The mechanism makes "tickets" available for sale to participants, who spend an endowment of Francs (our synthetic experimental currency) on tickets and allocate them across the buckets. At the opening of each application all participants are given a fixed budget of 500 Francs. The Francs cannot be transferred among participants, used in other applications, or assigned to buckets for another variable's IAM. As quality controls over the mechanism's operation, the IAM operates at a fixed time and only those invited are able to participate. The IAM program stores a wealth of data, including individual participant actions and time-stamps indicating when each of these actions took place.

The tickets for all buckets are priced the same and that price will move up at a pre-announced rate to ensure the mechanism closes in a reasonable time. For example, the opening price would be constant for fifteen minutes and then go up at a rate of one Franc per minute after that. These price changes discourage waiting until the last second to purchase, helping to offset individual incentives to hold back their private information and to improve their own information by herding on others' allocations. All participants are aware that their own information might be improved through seeing the purchases of others. They are also aware that their own information might be communicated by their own purchase of tickets. Inducing temporal discounting helps to mitigate these strategic incentives that otherwise hinder successful information aggregation. The price increase is constant but sufficiently substantial that by 40 minutes into the exercise the ticket prices are so high that the budget has little purchasing power.

Throughout the operation of the mechanism, participants have a continuously available record of the number of tickets that are currently placed in each of the buckets. At each instant during the application as well as at its termination, the placements of all tickets in all buckets are known. The individual participant also knows the proportion of tickets he or she holds in each bucket, which is particularly important because these proportions are the foundations for incentives. When the actual winning bucket becomes known those holding tickets in that bucket are given a part of a grand prize equal to the proportion of the winning bucket tickets that he or she holds. If participant n holds $z\%$ of the tickets sold for the winning bucket then participant n gets $z\%$ of the incentive prize. For example, if the incentive prize was $10,000 and the individual held 10% of the tickets sold for that bucket then the payment to participant n would be $1,000.[8]

Participants depend on the nature of the forecasting exercise. For forecasts that have significant influence on financial performance, only insiders, those with access to limited financially relevant information, are permitted; forecasts that are not considered material to earnings reports may include a wider group. Typically, the forecasters are insiders with direct access to the most the information relevant to the forecasting problem, either directly involved in management or sales. Data already available (to the insiders), including current signals and historical results, are packaged for all participants to study in preparation for the IAM exercise, establishing a base of relevant information to provide an underlying distribution of common knowledge. As such, it is important to synchronize the start time so that those individuals with appropriate information could participate.

A typical IAM exercise involves forecasting for the current quarter plus the three upcoming quarters. The exercise takes place once a month and requires on the order of 30 minutes. Each participant is given a separate Franc budget for each item they forecast. All budgets are the same size and the budgets are not fungible across the items forecast. The number of participants varies from ten to twenty-five and each operates from a secure computer located wherever the participant happened to be located, home, office, traveling, etc. Typically the users are anonymous within the mechanism: both the list of participants and the winners are secret. Of course, the total of tickets purchased in each bucket of each forecast is public and known in real time as the tickets are purchased. Interestingly, anonymity can be at odds with culture/tradition of openness and cooperation in matters of corporate decisions and that was the case with this application but the resistance dissolved with experience and success.

From an organizational perspective, Intel uses the IAM as an input to its official forecast. As such, the horizon h IAM forecast is released shortly before the horizon $h-1$ official forecasst. While the timing does not affect any analysis in evaluating the internal consistency

[8] The use of incentives inside a business reflected a belief and experience for experiments that incentives are central to the successful operations of information aggregation mechanisms. The performance of a mechanism without incentives (cheap talk) is explored by Bernnouri et al. (2011) as are the success of different measures of information aggregation.

of each forecast, it provides an important consideration in comparing forecasts.

4 Testing Rational Expectations and Information Aggregation

In the IAM, participants are incentivized to reveal their beliefs, or probabilities, that future unit sales will be realized within different ranges of values (the "buckets"). For each bucket x_k in the set of K buckets, $x_1, \ldots, x_K$, let $\hat{g}_{i,t|t-h}(x_k)$ denote the proportion of tickets in the mechanism at time $t-h$ participants place in bucket k. We then define the cumulative conditional distribution $\hat{G}_{i,t|t-h}(y) = \sum_{k=1}^{max\{\kappa|x_\kappa \leq y\}} \hat{g}_{i,t|t-h}(x_k)$ as representing the aggregated IAM forecast of the conditional distribution for $Y_{i,t}$ given the information available at time $t-h$. In this section, we test whether these manifested beliefs are rational, in the sense that these distributions coincide with the distribution of realized sales.

Mechanically, this exercise is complicated by the effect of conditioning information that ought to be incorporated in the mechanism's distributions at different horizons. For each product-period, let $Y_{i,t}$ denote the actual realization of unit sales from an unconditional distribution $F_{i,t}$. Under the null hypothesis that beliefs are correct, the actual outcome would be distributed according to $G_{(i,t|t-h)}$, the distribution of elicited beliefs conditional on information up to $t-h$. To operationalize the test we transform the realized outcome $Y_{(i,t)}$ into its corresponding quantile in the conditional belief distribution:

$$\hat{U}_{i,t,h} \equiv \hat{G}_{i,t|t-h}(Y_{i,t}) \sim_{H_0} U[0,1] \tag{1}$$

Accordingly, we can simply use a Kolmogorov-Smirnov test to evaluate whether we can reject that our sample of $\left\{\hat{U}_{i,t,h}\right\}_{t=1}^{T}$ are uniformly-distributed i.i.d. draws. By analyzing the conditional quantile, such a test is robust to heterogeneity in the distributions across products, time, and the information available.[9]

Figure 1 presents the empirical distribution for $\hat{U}_{i,t,h}$ plotted against the uniform distri-

[9] While robust, note that various features of our data, especially the panel structure coupled with multiple horizons, induce correlation across draws. As such, the p-Values of the Kolmogorov-Smirnov test are likely to be distorted with a downward bias. When the number of degrees of freedom for the Kolmogorov-Smirnov test is reduced by a factor of 4, the p-Value increases to 6.7%. While this correction is not valid, it does indicate a lack of robutstness for the results rejecting rational expectations. Unfortunately, allowing for correlated sampling structures in the Kolmogorov-Smirnov test is an intractable problem beyond our scope.

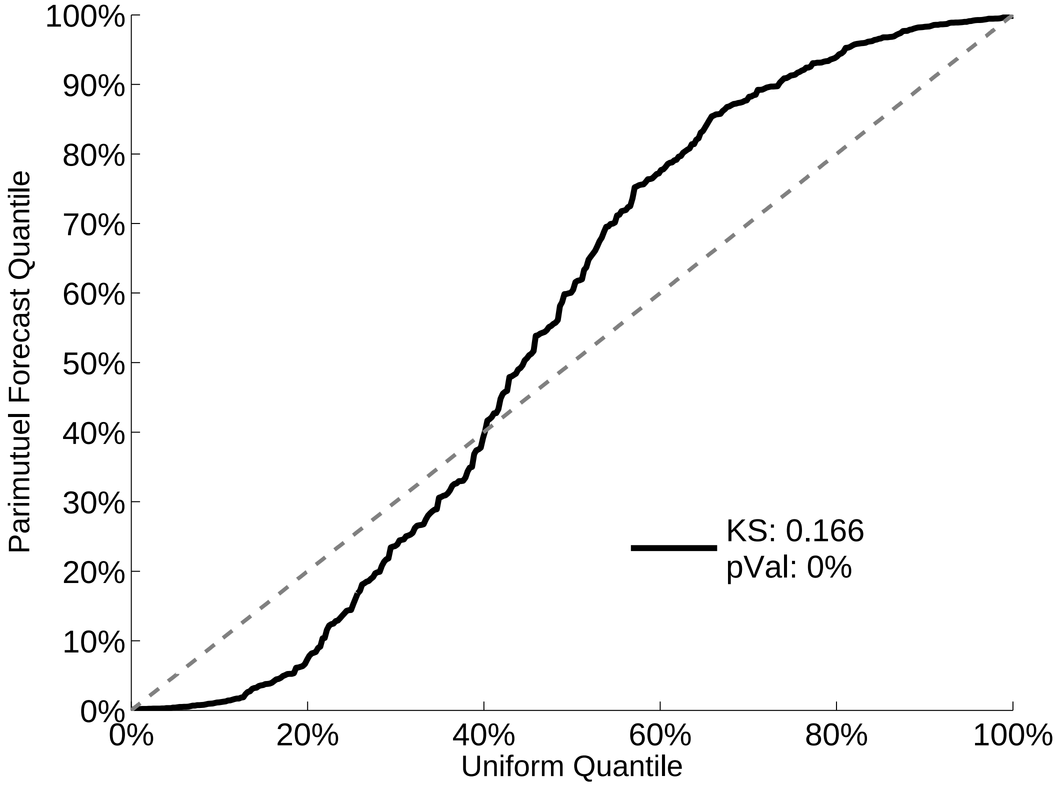


Figure 1: Quantile Plots for the Information Aggregation Mechanism
This figure presents the distribution of realized quantiles from the information aggregation mechanism in equation 1 against the theoretically accurate uniform distribution. The KS p-Value is reports the result for a Kolmogorov-Smirnov test of equality of the distributions.

bution quantiles pooled across all products, periods, and forecast horizons. Visual inspection indicates two features the mechanism's distribution appears to distort in the realized distribution over outcomes. First, the S-shape of the graph indicates a "reverse favorite-longshot bias"; that is, the beliefs from the mechanism understate the likelihood of extreme outcomes. Second, the quantiles for the mechanism's distribution appear slightly to the left of the uniform, indicating the IAM distribution is a bit conservative at the median. Because of these two distortions, the Kolmogorov-Smirnoff test rejects the null hypothesis that the empirical distribution of outcomes matches the IAM's forecasted distribution for outcomes.

Given that we observe conditional forecast distributions at different horizons, we can empirically evaluate how improved information affects the reverse FLB. That is, we'd like to test if the forecast distribution accuracy improves as the forecast horizon shortens. To that end, instead of pooling the $\hat{U}_{i,t,h}$ across horizons, we can perform the Kolmogorov Smirnov test from equation 1 for each horizon sub-sample.[10] Figure 2 plots the quantiles of the forecast quantile distribution against the uniform distribution at horizons 1, 3, 5, 7, and 9. Notably, the plots appear to approach the 45 degree line as the horizon drops, as indicated by the pattern of Kolmogorov-Smirnov test statistics and p-Values. Hence, the reverse favorite-longshot bias described previously disappears as the forecasting horizon shrinks.

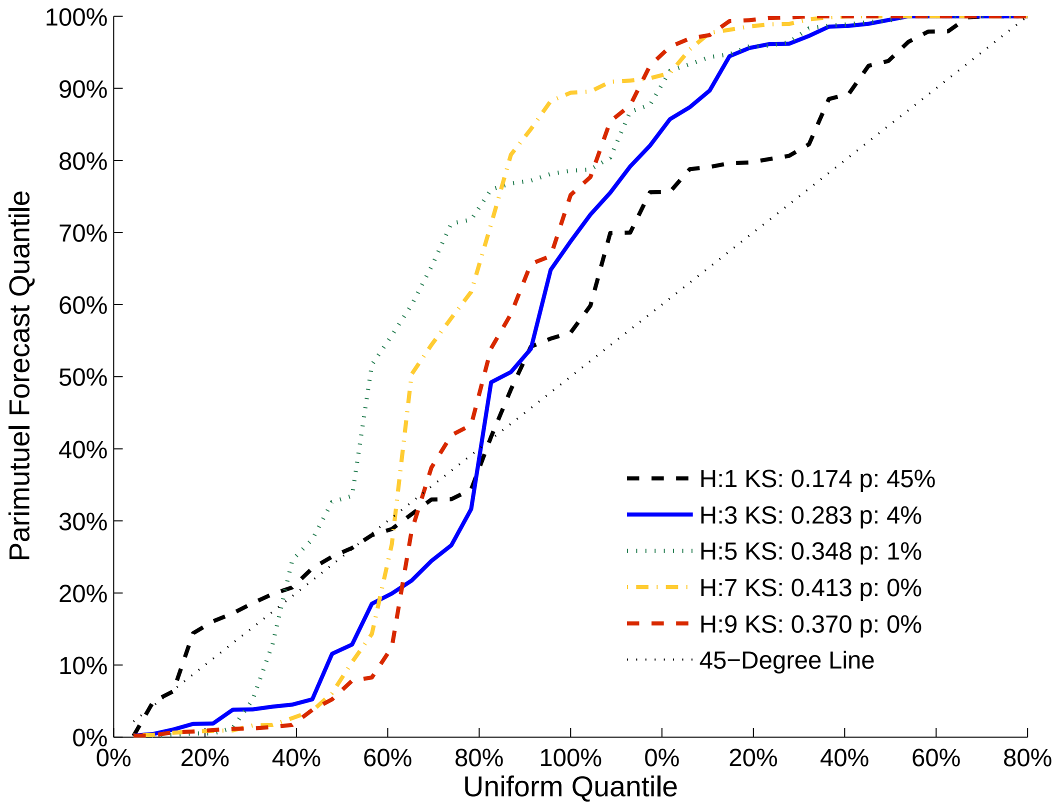


Figure 2: Quantile Plots for the Information Aggregation Mechanism by Horizon
This figure presents the distribution of realized quantiles from the information aggregation mechanism against the theoretical uniform distribution for horizon subsamples of the IAM. The KS p-Value reports the result for a Kolmogorov-Smirnov test of equality of the distributions.

Our finding that, at least for long horizons, the beliefs elicited from the mechanism systematically understate tail probability events, is related to phenomena which have been much studied in the literature on betting markets. Specifically, the "favorite-longshot bias (FLB)"

[10]Considering sub-samples also helps mitigate the effect of correlation in the sampling process for $\hat{U}_{i,t,h}$. In particular, one would expect a significant correlation in the conditional quantiles across horizons for a given product and period. Performing the test horizon-by-horizon prevents this distortion from affecting inference at the cost of weaker power due to fewer degrees of freedom. However, it doesn't fully control for the temporal autocorrelation in $\hat{U}_{i,t,h}$ at fixed horizon and across products for a fixed time period.

is an oft-reported empirical property in studies on betting markets.[11] In data patterns characterized by the FLB, the parimutuel odds on high probability events understate the realized probabilities (e.g., the parimutuel odds on a horse "favored" to win the race understates the true odds of that horse winning). By contrast, we find a "reverse favorite-longshot bias," in which elicited beliefs understate the realized probabilities for *low-probability* (tail) events.

A number of explanations have been proposed for the FLB, including probabilistic misperceptions, risk preferences, belief heterogeneity, and information incentives (see Ali (1977)).[12] The fact that the reverse-FLB disappears as the forecast horizon narrows (as in Figure 2) casts doubt on the risk explanation, because these preferences should not change with the forecast horizon. If risk aversion is driving the reverse-FLB, then the extent of the reverse-FLB would be primarily driven by the likelihood of the outcome itself. Importantly, the bias is driven by the likelihood of the event, rather than the uncertainty in the forecast distribution. That is, risk aversion would cause the likelihood of low-probability tail events to be understated in the forecast distribution, regardless of the forecast horizon – this is not what we see.

Ottaviani and Sørensen (2010) propose an alternative model of biases in parimutuel betting systems based on strategic information revelation rather than misperceptions or risk preferences. In this model, parimutuel betting participants are partially informed about the realization of a random variable but aggregate uncertainty about the outcome persists. Rational behavior in their model allows for a FLB or a reverse-FLB in the aggregated distribution depending on the ratio of privately-held information to noise in the forecast variable. Specifically, the reverse-FLB arises when information is very diffuse. To see the intuition, consider the case of Lotto, a purely random parimutuel market. Since each number has an equal probabaility of being a winning number, any "favorites" which arise during the betting process must underpay, and "longshots" must overpay: that is, a systematic underweighting of low-probability events arises in the parimutuel odds.

This explanation implies that at long horizons, the noisiness of the information leads to a reverse-FLB, which goes away as the horizon narrows and the quality of information improves: this is what we see in Figure 2. Such an explanation is partly corroborated by

[11] There are four full chapters dedicated to its review alone in the *Handbook of Sports and Lottery Markets* (Hausch and Ziemba (2008a)).

[12] For instance, Snowberg and Wolfers (2010) compare the relative likelihood of risk preferences and probabilistic misperception in betting markets, finding probabilistic misperceptions to be relatively more likely, but are silent on the role of strategic considerations. Gandhi and Serrano-Padial (2012) show that belief heterogeneity among racetrack bettors can also induce a longshot bias in prices.

some limited analyst-level data which was available to us. In one sample, we found that transactions for instances of the mechanism corresponding to longer-horizon forecasts took place about 25% later than in shorter-horizon sessions. This willingness of participants to delay their transactions may reflect the noisiness of participants' information at distant horizons, and hence their willingness to sacrifice costs to learn the information revealed within the IAM by earlier participants.

5 Testing Forecast Rationality

Having characterized the IAM's ability to aggregate information within the firm, we now ask how the mechanism can improve internal forecasts and decision-making. Many decisions, such as communicating expectations to Wall Street analysts or setting sales incentive quotas, require management to determine a point forecast for unit sales. In these settings, we are less interested in the full conditional distribution of sales than in a central measure of the distribution.

In this section, we evaluate the degree to which the official forecast and forecasts derived from the mechanism are internally consistent as rational forecasts. In addition to the canonical Mincer and Zarnowitz (1969) regressions testing forecast rationality at each individual horizon, we can exploit our forecast panel to implement more powerful multiple-horizon tests of forecast revision proposed by Patton and Timmermann (2012). We find little evidence that the available forecasts fail to satisfy rationality overall, suggesting only that revisions in the period before unit sales are realized appear to be understated.[13]

5.1 Forecast Data

We observe data from 2007 through 2010 across five major product lines. These forecasts are made in a dynamic business envornment where expectations rapidly shifted from growth to stagnation. With the Lehman brothers crash occuring half-way through the sample, the forecasting exercise is particularly challenging at longer horizons. For proprietary reasons, Intel has requested we mask the actual values of units sold as well as the names of the

[13] Unfortunately, this analysis suffers from the major limitation that findings consistent with rationality are largely negative results. Importantly, we fail to reject the null hypothesis that the forecast is not distorted, rather than rejecting the null that the forecast is distorted. Nonetheless, the evidence provided in favor of well-behaved forecasts is encouraging. For more discussion of the methodology of forecast evaluation, we refer interested readers to the excellent survey by West (2006).

products themselves. As all of our comparative analyses are insensitive to the numeraire, this masking has no effect on the results.

To consolidate the distributional information from the mechanism into point forecasts, we propose two natural point-forecast estimates. The Mean forecast is taken as the expected value of the outcome under the forecasting distribution. The Mode forecast is taken as the outcome with maximal probability under the forecasting distribution. Due to the buckets in the mechanism, these forecasts are effectively interval-censored. To address this censoring, we take the mid-point of the bucket as representing the value for all forecast mass placed within that bucket.[14] Recalling the definition of actual unit sales for product line i in quarter t by $Y_{i,t}$, we refer to the official, mean, and mode forecasts at horizon $h \in \{1, 2, \ldots, 0\}$ months by $\hat{Y}_{i,t|t-h}^{(\text{Official})}$, $\hat{Y}_{i,t|t-h}^{(\text{Mean})}$, and $\hat{Y}_{i,t|t-h}^{(\text{Mode})}$ respectively. For each $k \in \{\text{Official, Mean, Mode}\}$, we can then define the forecast error as $e_{i,t|t-h}^{(k)}$ from which we can compute forecast performance statistics such as the Root Mean Square Forecast Error (RMSFE).

Table 1 reports summary statistics for the point forecasts and actual unit sales, including a break down by product lines. Realized quarterly sales, averaging 21 million in the full sample, have a standard deviation over \$7 million, indicating a highly variable forecasting environment. On average, the forecasts slightly overstate average revenue, with the Mean and Mode average forecast improving upon the Official forecast bias by less than 100,000. Still, the IAM mean and mode forecasts deliver a root mean square error almost 10% lower than that of the the Official forecast. Within individual product lines, the RMSFE improvement is even more significant, ranging from 21% to 37%. By comparing the RMSFE with the standard deviation of sales within each product line, though, we see the forecasts in general perform quite poorly relative to the ex-post sample average. This poor performance may call into question the validity of any of the forecast mechanisms, a topic we consider empirically in Section 5.2.

Direct comparison of the forecasts is somewhat complicated by the timing of the IAM vis-a-vis the official forecasts. Since the mechanism takes place with a significant lag after the official forecast is released, the horizon t official forecast is most comparable to the horizon $t+1$ IAM forecast. Operationally, this implementation allows the analysts preparing the official forecast to observe the mechanism's distribution before announcing their forecast. When the last (most inaccurate) horizon is dropped from the official forecast and the nearest (most accurate) horizon is dropped from the IAM forecasts, the differential RMSFE is negligible

[14] The first and last buckets representing ranges $[0, x_1)$ and $[x_{K-1}, \infty)$ are assigned values x_1, and x_{K-1}, respectively.

Table 1: Summary Statistics for Forecast Data

This table presents summary statistics characterizing the average and standard deviation of unit sales, the target variable to be forecast, and the different forecasts available to decision makers. Due to variation in timing of the horizon forecasts relative to realized sales, the full-sample Root Mean Square Forecast Error is not directly comparable between the official and IAM forecasts. As such, the official RMSFE is also reported excluding the least-informed (9 month) horizon forecast and the IAM RMSFE is also reported excluding the most-informed (1 month) horizon forecast. The columns report results broken down by the product line for which forecasts are obtained to characterize heterogeneity.

	Full	Business Product Line				
	Sample	Disti 1	Disti 2	Direct 1	Direct 2	Direct 3
First Sales Period Quarter	2007Q1	2008Q1	2007Q2	2008Q3	2007Q1	2009Q1
Last Sales Period Quarter	2010Q1	2010Q1	2010Q1	2010Q1	2010Q1	2010Q1
Number of Forecast-Period-Product Obs	414	81	108	63	117	45
Number of Sales-Product Periods	46	9	12	7	13	5
Number of Forecasts per Product-Period	9	9	9	9	9	9
Average Unit Sales	2.95	1.94	1.97	3.77	3.69	4.00
Std Dev of Unit Sales	1.02	0.21	0.16	0.36	0.74	0.56
Average Official Forecast	2.99	2.01	2.00	3.84	3.71	4.07
Std Dev of Official Forecasts	1.03	0.18	0.18	0.43	0.79	0.39
Official Root Mean Square Fcst Error	0.48	0.26	0.22	0.54	0.63	0.65
Official RMSE excluding 9th Horizon	0.45	0.25	0.20	0.51	0.60	0.60
Average Parimutuel Mean Forecast	2.98	1.99	1.98	3.88	3.71	4.03
Std Dev of Parimutuel Mean Forecasts	1.04	0.18	0.16	0.41	0.78	0.39
Parimutuel Mean Root Mean Square Fcst Error	0.43	0.25	0.18	0.51	0.58	0.54
Mean RMSE excluding 1st Horizon	0.46	0.26	0.19	0.54	0.61	0.57
Average Parimutuel Mode Forecast	2.98	1.99	1.98	3.89	3.70	4.04
Std Dev of Parimutuel Mode Forecasts	1.04	0.19	0.16	0.42	0.78	0.41
Parimutuel Mode Root Mean Square Fcst Error	0.43	0.26	0.18	0.52	0.59	0.52
Mean RMSE excluding 1st Horizon	0.46	0.27	0.19	0.55	0.62	0.55

in aggregate. However, we will evaluate this hypothesis more formally in Section 6.

While there is some apparent heterogeneity across product lines, the magnitudes of realized unit sales are of the same order. Notably, though, all the tests we use are robust to product-line heterogeneity subject to appropriate clustering for computing standard errors. This robustness allows us to enhance the power of our tests by pooling the data across product lines, which is particularly helpful given the limited sample size we have available. In our statistical tests, we cluster our observations to correct for any autocorrelation introduced by the pooling and also analyze subsamples to ensure robustness of the findings from the pooled sample.

5.2 Testing Individual Forecasts

At each horizon, we can evaluate the degree to which each of the available forecast mechanisms is "rational" given observed unit sales by testing whether a transformation of that forecast could reduce loss. We first consider simple transformations to the location and scale of the forecast using the standard Mincer and Zarnowitz (1969) (henceforth, MZ) regression tests for forecast rationality. For forecast mechanism k, we consider the following regression:

$$Y_{i,t} = \alpha_k + \beta_k \hat{Y}^{(k)}_{i,t|t-h} + \epsilon_{i,t|t-h,k} \tag{2}$$

We then test the joint null hypothesis that $\alpha_k = 0$ (the forecast location is unbiased) and $\beta_k = 1$ (the forecast scale is unbiased) using the standard F-Test with appropriately robust standard errors.[15] For example, if α_k is significantly positive, the forecast systematically understates expected sales. Similarly, if β_k is significantly less than one, the forecast is systematically too far away from the unconditional average, essentially overstating the variability in sales implied by the available information.

In Table 2, we report the results of the MZ regressions and F-tests at each horizon from the official and IAM forecast models. At short horizons and especially one month from the end of the quarter, each of the forecasts are remarkably consistent with theory, with the estimated model nearly exactly matching the restricted model. At longer horizons, there is some evidence of forecast distortion. Though this distortion is not statistically significant, it does appear that long-horizon forecasts could be improved by shifting them towards the unconditional average forecast.

The MZ regression tests can be pooled across horizons after making a multiple-comparisons adjustment for critical values. We use a Bonferroni correction for these nine tests, which is quite conservative in contexts where the tests themselves are based on correlated samples across experiments, but simple and easy to implement. This pooled regression specification gives an overall perspective of forecast performance that can yield improved power. However, not being able to reject rational forecasting in any of the individual subsamples, it's no surprise that the pooled test estimates are also consistent with the restrictions implied by theory.

[15] Optimal forecast errors for an h step ahead forecast are generally autocorrelated at up to $h-1$ lags, requiring HAC standard errors for regression 2. In addition, we cluster observations for each quarter of unit sales and across product lines to ensure robustness to common temporal shocks and product-level heterogeneity.

Table 2: Mincer-Zarnowitz Tests for Point Forecast Accuracy

This table shows parameter estimates and clustered (across products) standard errors for the Mincer-Zarnowitz regression (Equation 2) for the official and IAM point forecasts at each horizon (Panel A) and pooled across horizons (Panel B). We also report the F-Statistic and p-Value for testing the joint null hypothesis of zero intercept with unit coefficient. The p-Values reported in Panel B test the hypotheses that the intercept equals zero and the coefficient equals unity individually and jointly using an F-test and are Bonferroni-corrected for testing across multiple horizons. All standard deviations are robust to autocorrelation up to the forecast horizon and clustered by forecasting quarter and product line.

Panel A: Horizon-by-Horizon Mincer-Zarnowitz Tests

	Official Forecast			Mean Forecast			Modal Forecast		
	Alpha (StDev)	Beta (StDev)	F-Test (p-Val)	Alpha (StDev)	Beta (StDev)	F-Test (p-Val)	Alpha (StDev)	Beta (StDev)	F-Test (p-Val)
1 Month	0.01	1.00	0.02	0.03	0.99	0.57	0.01	1.00	0.30
	(0.05)	(0.02)	(98%)	(0.02)	(0.01)	(57%)	(0.03)	(0.01)	(75%)
2 Months	0.12	0.96	0.47	0.08	0.97	0.41	0.08	0.97	0.57
	(0.16)	(0.07)	(63%)	(0.10)	(0.05)	(66%)	(0.09)	(0.04)	(57%)
3 Months	0.16	0.95	0.60	0.13	0.96	0.49	0.15	0.95	0.77
	(0.15)	(0.08)	(55%)	(0.12)	(0.06)	(61%)	(0.10)	(0.05)	(47%)
4 Months	0.17	0.94	0.48	0.15	0.95	0.49	0.18	0.94	0.71
	(0.25)	(0.12)	(62%)	(0.18)	(0.10)	(62%)	(0.15)	(0.08)	(50%)
5 Months	0.31	0.88	1.46	0.23	0.92	0.95	0.27	0.91	1.22
	(0.30)	(0.15)	(24%)	(0.24)	(0.12)	(40%)	(0.25)	(0.12)	(30%)
6 Months	0.33	0.87	1.47	0.31	0.88	1.45	0.31	0.88	1.82
	(0.31)	(0.16)	(24%)	(0.27)	(0.14)	(24%)	(0.25)	(0.13)	(17%)
7 Months	0.42	0.84	2.48	0.37	0.86	2.17	0.41	0.85	2.04
	(0.29)	(0.16)	(10%)	(0.29)	(0.15)	(13%)	(0.30)	(0.16)	(14%)
8 Months	0.49	0.80	4.15	0.42	0.83	3.20	0.43	0.83	3.32
	(0.26)	(0.15)	(2%)	(0.27)	(0.15)	(5%)	(0.26)	(0.14)	(5%)
9 Months	0.51	0.79	4.48	0.50	0.80	4.05	0.52	0.80	3.75
	(0.24)	(0.14)	(2%)	(0.24)	(0.14)	(2%)	(0.25)	(0.15)	(3%)
Num of Obs	46								

Panel B: Joint MZ Tests

	Official Forecast			Mean Forecast			Modal Forecast		
	Alpha	Beta	F-Test*	Alpha	Beta	F-Test*	Alpha	Beta	F-Test*
Pooled Sample	0.29	0.89	14.69	0.25	0.90	13.34	0.27	0.90	14.05
	(0.17)	(0.09)	(0%)	(0.15)	(0.08)	(0%)	(0.15)	(0.07)	(0%)
Num of Obs	414								

*Bonferonni corrected p-Values

5.3 Multiple-Horizon Tests of Forecast Revisions

With forecasts available at multiple horizons, we can impose a stronger form of rationality. In addition to testing whether the forecasts themselves could be improved by a simple translation of location and scale, we can also evaluate whether the revisions made at each horizon could be similarly improved. These additional restrictions admit much more powerful tests of forecast rationality that indicate some potential inefficiencies in the revisions to forecasts from both the official and the IAM. In particular, the last revision made to the forecast one month prior to the realization of unit sales tends to be understated and overall forecasting could be improved by amplifying that revision.

Define the forecast revision from a long horizon (h_L) to a short horizon (h_S) as $d^{(k)}_{i,t|h_L \to h_S} =$

$\hat{Y}^{(k)}_{i,t|t-h_S} - \hat{Y}^{(k)}_{i,t|t-h_L}$ for each $k \in \{\text{Official, Mean, Mode}\}$, product, and revenue period. Then letting H denote the longest horizon at which a forecast is available, the natural extension to the MZ regression is the "optimal revision" regression proposed by Patton and Timmermann (2012):

$$Y_{i,t} = \alpha_k + \beta_{k,H}\hat{Y}_{i,t|t-h_L} + \sum_{j=1}^{H-1} \beta_{k,j} d^{(k)}_{i,t|h_L \to h_S} + \epsilon_{k,i,t} \tag{3}$$

As in the standard MZ regression, if the forecast is rational at all horizons, then it will be unbiased across horizons (i.e., α_k will be zero). Additionally, if each period's forecast revision has the correct scale given the amount of information revealed between each of the long and short horizons. As such, under rational forecasting, the regression 3 will satisfy the joint restrictions:

$$H_0 : \alpha_k = 0 \cap \beta_{k,1} = 1 \cap \cdots \cap \beta_{k,H} = 1 \tag{4}$$

We test these restrictions using an F-test, with the results appearing in Table 3. Here, we see that most of the forecast revisions (the coefficients for periods greater than one month) are all statistically near unity. However, the last forecast revision at a one month horizon seems to be typically understated, particularly for the Official forecast but also for the Mean forecast, both of which reject the MZ restriction at almost any size. As a result, we can reject the null hypothesis that the forecasts are perfectly rational in their revisions with essentially a zero p-value.[16]

In addition to the restrictions implied by the hypothesis 4, Patton and Timmermann (2012) present a number of additional restrictions the multiple-horizon setting imposes on second moments for forecasts, revisions, errors, and realized unit sales. In our implementation of these tests, we found no further evidence of forecast breakdown, so we refer interested readers to the results and further discussion in Appendix A1.

[16] One potential explanation for this systematic "understatement" of the final forecast revision may be found in Brandenburger and Polak (1996), who show that managers interested in maximizing the share price of their company may take actions which contradict their private information, but are in line with the prior expectations of market watchers. In our setting, such an effect can arise if IAM participants do not want to appear to be overly contradicting their earlier forecasts.

Table 3: Mincer-Zarnowitz Optimal Forecast Revision Tests

This table shows the results of the Mincer-Zarnowitz Optimal Revision regression 3 for the Official and IAM Forecasts. Standard Errors are HAC and clustered for individual products and revenue periods.

	Official Forecast		Mean Forecast		Mode Forecast	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
Intercept	(0.01)	0.05	0.04	0.03	0.02	0.05
1 Month	1.50**	0.10	1.31**	0.06	1.16	0.10
2 Months	1.05	0.30	0.89	0.12	1.06	0.15
3 Months	0.87	0.15	0.80	0.14	0.84	0.13
4 Months	0.99	0.14	1.01	0.10	0.94	0.07
5 Months	0.92	0.14	1.15*	0.07	1.19*	0.10
6 Months	1.07	0.15	0.92	0.16	0.90	0.11
7 Months	0.75	0.14	0.96	0.08	1.02	0.04
8 Months	1.19	0.14	0.94	0.11	0.94	0.09
9 Months	1.00	0.02	0.99	0.01	1.00	0.02
F-Test Statistic	6.15**		15.68**		2.67**	
p-Value	0%		0%		1%	
Number of Obs	46		46		46	

*,** denote significance at the 5% and 1% levels, respectively

6 Comparing the Information Aggregation Mechanism to Official Forecasts

Having established that the IAM and official forecast are both rational forecasts in the previous section, we now evaluate their relative performance. We begin by directly comparing the loss due to forecast errors, finding the IAM forecasts outperform the official forecast under quadratic loss in those settings where high-quality individual information may be dispersedly held. For example, the IAM performs particularly well at short forecast horizons and in direct distribution channels where internal participants are likely to have useful information for the forecasting exercise. In contrast, at long horizons and in channel distributor based product lines, the official forecast based on more structured models proves to be more informative.

6.1 Predictive Accuracy with Quadratic Loss

Diebold and Mariano (1995), henceforth DM, tests provide the benchmark for directly comparing the predictive accuracy of two forecasts under a variety of possible loss functions. While we later consider a broader class of asymmetric loss functions in section 7.2, here we simply define forecast loss as the square error of the forecast:

$$l_{i,t|t-h}^{(k)} = \left(\hat{Y}_{i,t|t-h}^{(k)} - Y_{i,t}\right)^2 \tag{5}$$

We can then compare the loss between two corresponding forecasts j and k:

$$\delta_{i,t,h}^{(j,k)} = l_{i,t|t-h}^{(j)} - l_{i,t|t-h}^{(k)} \tag{6}$$

The DM test statistic corresponds to the t-statistic for the average $\delta_{i,t,h}^{(j,k)}$, using a robust estimator of the variance allowing for auto-correlation of loss differentials within product lines and clustering for each revenue period. While DM's initial derivation of the test establishes its asymptotically normality, Harvey et al. (1997) show that Student's t distribution better controls for size.

We present the results of these tests in Table 4, comparing the loss of the horizon h IAM forecast with the horizon $h-1$ official forecast. Given the differential timing of the official forecast release and the mechanism, this treatment cedes a slight information advantage to the official forecast, which is always released *after* the mechanism has concluded. Despite this informational advantage, the full sample results indicate the point forecasts taken from the mechanism nonetheless deliver lower square loss than the official forecasts. This outperformance is especially surprising given that the offical forecasters know the IAM distribution *before* releasing their forecast. That is, the analysts' deviation from the IAM forecast actually worsens the forecast error.

Using DM tests to evaluate the mechanism's performance in subsamples, we find the IAM perform particularly well in exactly those contexts where individuals within the organization are likely to have disparate information. Conversely, the official forecast performs well in those settings that are conducive to high-level modeling and analysis based on historical and macroeconomic trends.

The root mean forecast improvement is monotonically declining in forecast horizon, with the official forecast outperforming the information aggregation mechanism at the 8 and 9 month horizons. The official forecast is also more likely to incorporate higher-level information regarding macroeconomic conditions that is especially relevant to long-horizon forecasting. Given that individual sales representatives and partner managers are unlikely to have materially relevant information about unit sales at these horizons, the official forecast's outperformance is to be expected here.

Table 4: Comparing Forecast Loss Across Mechanisms

This table presents Diebold-Mariano tests comparing the point forecasts from the official forecast and the mean and mode information aggregation mechanism forecast. Panel A uses the full sample of all forecasts and horizons, with Panels B and C reporting results for horizon and product subsamples. The Root Mean Δ Square Error reports the square root of the absolute average difference in the square error for the official and IAM forecasts, signed negatively for cases where the official forecast outperforms the IAM. The Outperformance Frequency captures the frequency with which the IAM forecast was more accurate than the official forecast. The DM-Statistic and p-Value report the Diebold-Mariano test statistic and p-Value using standard errors robust to autocorrelation up to the maximum horizon included in the sample, clustered by period and product.

Panel A: Pairwise Diebold Mariano Tests

	Mean Forecast				Mode Forecast				
	Root Mean Δ Square Error	Outperf Freq	DM Stat	p-Val	Root Mean Δ Square Error	Outperf Freq	DM Stat	p-Val	Num of Obs
Overall	0.45	61%	(2.20)*	3%	0.45	60%	(2.13)*	3%	368

Panel B: Pairwise Diebold Mariano Tests by Horizon

	Mean Forecast				Mode Forecast				
Horizon	Root Mean Δ Square Error	Outperf Freq	DM Stat	p-Val	Root Mean Δ Square Error	Outperf Freq	DM Stat	p-Val	Num of Obs
2 Months	0.69	78%	(2.18)*	3%	0.70	78%	(2.19)*	3%	46
3 Months	0.66	80%	(1.90)	6%	0.66	76%	(1.89)	6%	
4 Months	0.63	80%	(1.68)	10%	0.63	76%	(1.68)	10%	
5 Months	0.49	70%	(1.57)	12%	0.49	63%	(1.57)	12%	
6 Months	0.42	54%	(1.42)	16%	0.43	61%	(1.44)	16%	
7 Months	0.19	48%	(0.55)	59%	0.10	50%	(0.16)	88%	
8 Months	(0.22)	39%	1.12	27%	(0.22)	41%	1.24	22%	
9 Months	(0.25)	35%	1.48	15%	(0.27)	35%	1.67	10%	

Panel C: Pairwise Diebold Mariano Tests by Product Line

	Mean Forecast				Mode Forecast				
Product Line	Root Mean Δ Square Error	Outperf Freq	DM Stat	p-Val	Root Mean Δ Square Error	Outperf Freq	DM Stat	p-Val	Num of Obs
Disti 1	(0.21)	36%	1.03	31%	(0.22)	35%	1.07	29%	72
Disti 2	(0.07)	50%	0.47	64%	(0.07)	50%	0.47	64%	96
Direct 1	0.70	75%	(1.56)	12%	0.69	77%	(1.54)	13%	56
Direct 2	0.63	75%	(1.65)	10%	0.62	71%	(1.54)	13%	104
Direct 3	0.54	73%	(1.02)	32%	0.57	78%	(1.07)	29%	40

Similarly, the official forecast outperforms the mechanism's expectation in products distributed through channel sales operations. These distribution networks correspond to smaller-sized transactions and, as such, are driven primarily by trends in consumer spending best analyzed in aggregate. Further, these sales operations generate only about 10% of total corporate profits. In contrast, for the direct channels that generate about 90% of firm profits, the IAM performs exemplary at aggregating the information held by individual agents.

On the whole, this analysis indicates the information aggregation mechanism performs well at forecasting in exactly those settings that economic theory indicates its usefulness. Overall, the mechanism outperforms the official forecast and it does so by dramatically reducing forecast loss in those settings where individually held information is likely to be particularly valuable.

6.2 Forecast Combination and Encompassing Tests

Instead of just choosing one forecast, we may wish to combine the information from the mechanism with the official forecasts into a single aggregated forecast. A robust literature considers optimal forecast combination, with the survey by Timmermann (2006) providing a good entry point. For the aggregated forecast to improve upon the IAM forecast, however, there must be some information in the official forecast that isn't already incorporated into the mechanism's expectations. As such, we'd like a test to make sure the mechanism's forecast doesn't encompass the official forecast before introducing it into a forecast combination exercise.

Following the approach of Fair and Shiller (1990), one way to implement such a test would be to buld on the MZ tests from section 5.2 for the IAM forecast by adding the Official forecast as an additional explanatory variable. Among the most basic results from the forecast combination exercise, it's straightforward to show that the optimal weights with which to form a linear combination of forecasts can be calculated using the following regression.

$$Y_{i,t} = \alpha + \omega_{IAM}\hat{Y}^{(IAM)}_{i,t|t-h} + \omega_{Official}\hat{Y}^{(Official)}_{i,t|t-h} \tag{7}$$

For instance, one way to evaluate whether the official forecast encompasses the Mean forecast is to test the null hypothesis that $\alpha = 0$, $\omega_{IAM} = 0$, and $\omega_{Official} = 1$. If we reject this null hypothesis using an F-test, then we can be confident that the mechanism's forecast has additional information beyond that which is contained in the Official forecast. Similarly, if we reject the null hypothesis that $\alpha = 0$, $\omega_{IAM} = 1$, and $\omega_{Official} = 0$ then we will be able to say that the Official forecast contains information beyond that which is available from the mechanism's point forecast. Slightly less restrictive forms of these tests can also be formulated by dropping the condition on α and still simpler tests can just evaluate the one-sided hypotheses that $\omega_{Official} \geq 0$ and $\omega_{IAM} \geq 0$.[17]

Table 5 summarizes the results of these regressions for the Mean and Mode forecast.

[17] These regression-based tests have been further generalized by Harvey et al. (1998), whose approach mirrors that of Diebold and Mariano (1995). We implement the Harvey, Leybourne, and Newbold tests in Appendix A2 with largely the same results presented here. We only present the results using the Fair and Shiller (1990) test here for brevity and ease of interpretation in relation to forecast combinations.

Table 5: Forecast Combination Regressions

This table presents estimates from the forecast combination regressions 7. Panel A uses the full sample of all forecasts and horizons, with Panels B and C reporting results for horizon and product subsamples. The test F(0, 1, 0) tests the hypothesis that $\alpha = 0$, $\omega_{IAM} = 1$, and $\omega_{Official} = 0$, similarly, F(0, 0, 1) tests $\alpha = 0$, $\omega_{IAM} = 0$, and $\omega_{Official} =$. The tests F(., 0, 1) and F(., 0, 1) test the analogous restrictions without the zero-intercept condition. All tests use standard errors robust to autocorrelation up to the maximum horizon included in the sample, clustered by period and product.

Panel A: Full-Sample Regressions

	Intercept St Dev	Parimutuel Forecast St Dev	Official Forecast St Dev	F(0, 1, 0) p-Value	F(0, 0, 1) p-Value	F(., 1, 0) p-Value	F(., 0, 1) p-Value	
Mean Fcst	0.26	0.76	0.14	12.80	144.82	17.63	215.78	368.00
	0.16	0.14	0.08	0%	0%	0%	0%	
Mode Fcst	0.27	0.75	0.15	14.08	145.06	19.67	216.13	368.00
	0.15	0.14	0.08	0%	0%	0%	0%	

Panel B: Mean Forecast by Horizon

Horizon	Intercept St Dev	Parimutuel Forecast St Dev	Official Forecast St Dev	F(0, 1, 0) p-Value	F(0, 0, 1) p-Value	F(., 1, 0) p-Value	F(., 0, 1) p-Value	
h = 1	0.06	0.95	0.03	0.40	143.88	0.57	215.82	46.00
	0.11	0.04	0.02	75%	0%	57%	0%	
h = 2	0.10	0.90	0.07	0.68	73.79	1.01	110.67	
	0.12	0.07	0.04	57%	0%	37%	0%	
h = 3	0.12	0.88	0.08	0.68	45.96	1.01	68.93	
	0.18	0.12	0.05	57%	0%	37%	0%	
h = 4	0.22	0.85	0.07	0.74	20.01	1.06	29.47	
	0.24	0.21	0.11	54%	0%	36%	0%	
h = 5	0.29	0.71	0.18	1.52	12.52	2.16	18.35	
	0.28	0.26	0.15	22%	0%	13%	0%	
h = 6	0.37	0.48	0.38	3.11	5.29	4.33	7.85	
	0.31	0.26	0.13	4%	0%	2%	0%	
h = 7	0.48	0.07	0.75	6.38	3.25	8.75	4.39	
	0.31	0.30	0.25	0%	3%	0%	2%	
h = 8	0.54	(0.03)	0.82	6.48	3.06	8.75	4.23	
	0.34	0.25	0.17	0%	4%	0%	2%	

Panel C: Mean Forecast by Product

Product	Intercept St Dev	Parimutuel Forecast St Dev	Official Forecast St Dev	F(0, 1, 0) p-Value	F(0, 0, 1) p-Value	F(., 1, 0) p-Value	F(., 0, 1) p-Value	
Disti 1	1.05	(0.23)	0.70	72.74	12.00	104.04	17.68	72
	0.13	0.10	0.13	0%	0%	0%	0%	
Disti 2	1.15	0.06	0.36	22.85	15.01	33.98	20.69	96
	0.29	0.26	0.18	0%	0%	0%	0%	
Direct 1	4.53	0.09	(0.29)	35.70	125.75	49.72	188.61	56
	0.28	0.09	0.06	0%	0%	0%	0%	
Direct 2	1.34	0.66	(0.02)	8.96	53.23	13.35	79.57	104
	0.80	0.23	0.09	0%	0%	0%	0%	
Direct 3	2.59	0.41	(0.06)	1.92	14.50	2.82	13.69	40
	3.34	0.47	0.49	14%	0%	7%	0%	

These regressions show that the optimal combination of these forecasts negatively weight the official forecast. The F-Tests indicate that we can reject the null hypothesis that the IAM forecast encompasses the official forecast, as the Official forecast contains information that can be used to moderate errors in the IAM forecast estimates. Indeed, even considering just the coefficient on the Official forecast, we can reject the minimally restrictive hypothesis that $\omega_{Official} \geq 0$.

The subsample results for Table 5 in panels B and C focus on just combining the mechanism's Mean forecast with the Official forecast at each horizon, though the results are similar although slightly weaker for the Mode forecast. The evidence in these subsamples is broadly consistent with the full-sample results. The significance of the results drops somewhat in the horizon-based subsamples, particularly at the longer horizons. Also, as in the Mincer-Zarnowitz regressions, the estimated coefficients for the product subsamples occasionally deviate widely from the values hypothesized by the encompassing tests.

7 Higher Moments in the IAM and Forecasting

The tests in the previous sections focus on the performance of point forecasts extracted from the mechanism's distribution of outcomes in comparison to the official forecast. However, this focus disregards much of the information available in the mechanism, specifically any higher moments from the distribution that characterizes uncertainty in the forecast itself. We now focus on this additional information and how it might be used to improve the official forecast either under quadratic or under asymmetric forecast loss functions. As would be expected under quadratic loss, we find that higher moments do not assist in forecasting unit sales themselves. Still, in settings a forecast that overstates realized sales is asymmetrically worse than a forecast that understates sales, we find that the IAM performs distinctly well.

7.1 Higher Moments in Forecasting Sales under Quadratic Loss

While the mechanism's mean and mode forecasts both perform well and encompass one another, we might ask whether higher moments of the forecast distribution could also improve forecast performance. Denoting the mth central moment from a given IAM forecast distribution by $\hat{\mu_{m,i,t|t-h}}$, we could form a hybrid forecast:

$$\hat{Y}_{i,t|t-h}^{Hybrid,M} = \theta_0^{(M)} + \sum_{m=1}^{M} \theta_0^{(M)} \hat{\mu}_{m,i,t|t-h} + \epsilon_{i,t|t-h}^{(M)} \tag{8}$$

Note that here, we are not considering the significance of the θ parameters, rather only the degree to which we can credibly improve our sales forecast using additional information from the IAM distribution. Consolidating this additional information requires specifying a model for doing so, an exercise that would typically be estimated using a pre-sample to fit a current period's forecast. To model this feature, we use a blocked cross-validation approach whereby we estimate the model using a "leave-one-t-out" approach.[18] Denoting these estimated parameters for the forecasting model as $\hat{\theta}_{0,-t}^{(M)}, \ldots, \hat{\theta}_{M,-t}^{(M)}$, we estimate the forecasts for period t across all horizons and all products as:

$$\tilde{Y}_{i,t|t-h}^{Hybrid,M} = \hat{\theta}_{0,-t}^{(M)} + \sum_{m=1}^{M} \hat{\theta}_{m,-t}^{(M)} \hat{\mu}_{m,i,t|t-h} + \epsilon_{i,t|t-h}^{(M)} \tag{9}$$

That done, we can then define the estimated forecast error for the hybrid forecasts and compare this to the IAM Mean's forecast errors. The results from this analysis are reported in Table 6.

Perhaps surprisingly, we find the hybrid forecast models almost uniformly underperform the mechanism's mean forecast. Though estimation error is well known to have an adverse impact on forecast combination, the IAM mean forecast delivers less error than any statistically fitted model in any subsample. Looking to higher moments from the distribution to refine the hybrid forecast improves the hybrid forecasts' performance slightly. In many cases, this effect is not statistically significant, but the cumulative evidence indicates that the forecast mean does contain most of the relevant information for estimating final unit sales.

[18] Specifically, for each quarter (e.g., 2009Q2) we estimate the regression 8 for $M = 1, 2, 3, 4$ using all forecast data except the data for that quarter (e.g., all data for revenue quarters up to 2009Q1 and after 2009Q3, inclusively). Arguably, this blocking structure may retain some autocorrelation in the error structures between the training and testing samples, a conservative treatment in the context of the results.

Table 6: Cross-Validation Tests of Forecasts with Higher Moments

This table characterizes the performance of linear forecasting models incorporating higher moments from the information aggregation mechanism's distribution characterized by equation 8. We use a leave-one-t-out approach to calculate cross-validation square loss and compare it with the loss from the IAM mean forecast. RMSE is the root mean square forecast error and CV RMSE is its cross-validation analog. Outperformance Frequency characterises the frequency with which the IAM mean forecast outperforms the fitted model. The DM Test p-Val reports the p-value for a Diebold-Mariano test that the IAM mean forecast has lower expected loss than the fitted model, computed using standard errors robust to autocorrelation up to the maximum horizon included in the sample, clustered by period and product.

Panel A: Full Sample Fitted Forecast vs Mean Forecast Performance

	Mean	Fitted Forecast: μ_1			Fitted Forecast: μ_1, μ_2			Fitted Forecast: μ_1, μ_2, μ_3			Fitted Forecast: $\mu_1, \mu_2, \mu_3, \mu_4$		
	Fcst	CV	Outperf	DM Test	CV	Outperf	DM Test	CV	Outperf	DM Test	CV	Outperf	DM Test
	RMSE	RMSE	Freq	p-Val	RMSE	Freq	p-Val	RMSE	Freq	p-Val	RMSE	Freq	p-Val
Overall	0.46	0.49	61%	3%	0.48	61%	25%	0.46	56%	96%	0.46	57%	86%

Panel B: Fitted Forecast vs Mean Forecast Performance by Horizon

	Mean	Fitted Forecast: μ_1			Fitted Forecast: μ_1, μ_2			Fitted Forecast: μ_1, μ_2, μ_3			Fitted Forecast: $\mu_1, \mu_2, \mu_3, \mu_4$		
	Fcst	CV	Outperf	DM Test	CV	Outperf	DM Test	CV	Outperf	DM Test	CV	Outperf	DM Test
Horizon	RMSE	RMSE	Freq	p-Val	RMSE	Freq	p-Val	RMSE	Freq	p-Val	RMSE	Freq	p-Val
1 Month	0.22	0.24	65%	8%	0.25	61%	6%	0.25	61%	5%	0.27	67%	7%
2 Months	0.30	0.33	61%	15%	0.33	59%	15%	0.33	63%	12%	0.37	70%	17%
3 Months	0.36	0.41	70%	17%	0.42	70%	9%	0.39	54%	7%	0.38	61%	18%
4 Months	0.43	0.48	65%	11%	0.46	50%	27%	0.44	52%	75%	0.52	61%	16%
5 Months	0.50	0.55	65%	10%	0.54	50%	16%	0.52	57%	39%	0.57	65%	13%
6 Months	0.53	0.59	63%	13%	0.58	72%	9%	0.58	70%	5%	0.57	61%	16%
7 Months	0.57	0.61	59%	30%	0.54	50%	62%	0.54	52%	49%	0.54	57%	54%
8 Months	0.61	0.64	61%	48%	0.59	59%	76%	0.64	63%	62%	0.69	61%	35%

Panel C: Fitted Forecast vs Mean Forecast Performance by Product Line

	Mean	Fitted Forecast: μ_1			Fitted Forecast: μ_1, μ_2			Fitted Forecast: μ_1, μ_2, μ_3			Fitted Forecast: $\mu_1, \mu_2, \mu_3, \mu_4$		
	Fcst	CV	Outperf	DM Test	CV	Outperf	DM Test	CV	Outperf	DM Test	CV	Outperf	DM Test
Product Line	RMSE	RMSE	Freq	p-Val	RMSE	Freq	p-Val	RMSE	Freq	p-Val	RMSE	Freq	p-Val
Disti 1	0.26	0.29	58%	28%	0.28	56%	56%	0.30	54%	29%	0.31	64%	19%
Disti 2	0.19	0.21	50%	31%	0.19	52%	52%	0.19	48%	86%	0.20	49%	75%
Direct 1	0.54	0.58	71%	17%	0.53	52%	94%	0.53	61%	89%	0.55	64%	65%
Direct 2	0.61	0.66	73%	10%	0.66	68%	20%	0.66	72%	35%	0.70	74%	10%
Direct 3	0.57	0.64	70%	33%	0.61	65%	36%	0.58	58%	88%	0.63	63%	35%

7.2 Optimal Forecasts under Asymmetric Loss

Forecast errors are often associated with asymmetric loss functions since failing to meet market expectations commonly has a much larger effect on a firm's valuation than exceeding those expectations. In this context, a conservative forecast that's biased downward to under-promise and over-deliver could be optimal for the organization even though it appears suboptimal from the perspective of symmetric loss functions. Analyzing models with asymmetric costs for forecast errors have long been a concern in evaluating forecast rationality, with references going back to Granger and Newbold (1986). Diebold and Mariano (1995) show their test is consistent under any well-behaved loss function as long as the loss function being known by the econometrician. Elliott et al. (2005, 2008) propose consistent tests of forecast rationality under unknown loss functions given instruments that characterize the forecaster's contemporaneous information set.

For our application, we simply modify the quadratic loss function to allow for asymmetric quadratic loss with a kink at zero imposing a higher cost to forecasts that are overly optimistic.

$$l\left(\hat{Y}, Y; \gamma\right) = \left(1 + \gamma 1\left\{\hat{Y} > Y\right\}\right)\left(\hat{Y} - Y\right)^2 \tag{10}$$

This specification restricts the set of loss functions somewhat, but has the benefit of being controlled by a single parameter. In this way, we can easily evaluate the forecasts in a reasonably flexible class of asymmetric loss functions by varying the downside loss aversion parameter, γ.

Given the proposed loss function, the IAM mean no longer gives an optimal forecast conditional the distribution of forecast uncertainty. Rather an optimal forecast would minimize expected loss under the mechanism's distribution:

$$\hat{Y}^*_{i,t|t-h}(\gamma) = \arg\min_{\hat{Y}} \int l\left(\hat{Y}, z; \gamma\right) d\hat{G}_{i,t|t-h}(z) \tag{11}$$

Being able to provide a forecast that optimizes asymmetric loss is an important feature available in the mechanism that cannot be readily captured in any other type of prediction market that delivers only point forecasts. Using the aggregated information from the mechanism, where we wouldn't expect participants to incorporate asymmetries into their bidding behavior, we get an unbiased estimate of the distribution $\hat{G}_{i,t|t-h}$. We can then account for asymmetries in the loss function to aggregate the information into a point forecast.

Using the asymmetric quadratic loss function defined in equation 10, we calculate the

Table 7: Diebold-Mariano Tests for Optimal Forecasts under Asymmetric Loss

This table presents Diebold-Mariano tests comparing the point forecasts from the official forecast and the IAM Mean forecast with the optimal forecast implied by the mechanism's distribution using asymmetric loss functions parameterized by γ as in equation 10. Panel A uses the full sample of all forecasts and horizons, with Panels B and C reporting results for horizon and product subsamples. The DM-Statistic and p-Value report the Diebold-Mariano test statistic and p-Value using standard errors robust to autocorrelation up to the maximum horizon included in the sample, clustered by period and product.

Panel A: Pairwise Diebold Mariano Tests

		Asymmetric Optimal vs Official				Asymmetric Optimal vs Mean				Num
	γ	0.1	1	10	100	0.1	1	10	100	of Obs
Overall	DM Statistic	(2.17)	(1.97)	(1.83)	(1.99)	(0.51)	(1.25)	(1.97)	(2.09)	368
	p-Value	3%	5%	7%	5%	61%	21%	5%	4%	

Panel B: Pairwise Diebold Mariano Tests by Horizon

		Asymmetric Optimal vs Official				Asymmetric Optimal vs Mean				Num
	γ	0.1	1	10	100	0.1	1	10	100	of Obs
1 Month	DM Statistic	(2.17)	(2.01)	(1.65)	(1.56)	(0.28)	(0.82)	(1.33)	(1.43)	46
	p-Value	4%	5%	11%	13%	78%	42%	19%	16%	
2 Months	DM Statistic	(1.89)	(1.79)	(1.55)	(1.50)	(0.02)	(0.49)	(1.04)	(1.13)	
	p-Value	7%	8%	13%	14%	98%	63%	31%	27%	
3 Months	DM Statistic	(1.67)	(1.59)	(1.42)	(1.42)	0.08	(0.46)	(1.02)	(1.09)	
	p-Value	10%	12%	16%	16%	94%	65%	31%	28%	
4 Months	DM Statistic	(1.57)	(1.51)	(1.45)	(1.49)	(0.13)	(0.69)	(1.13)	(1.19)	
	p-Value	12%	14%	15%	14%	89%	49%	27%	24%	
5 Months	DM Statistic	(1.41)	(1.35)	(1.35)	(1.46)	(0.34)	(0.83)	(1.19)	(1.24)	
	p-Value	17%	18%	18%	15%	74%	41%	24%	22%	
6 Months	DM Statistic	(0.51)	(0.39)	(0.75)	(1.21)	(0.60)	(0.99)	(1.26)	(1.29)	
	p-Value	61%	70%	46%	23%	55%	33%	21%	20%	
7 Months	DM Statistic	1.13	1.09	(0.42)	(1.49)	(0.50)	(0.95)	(1.29)	(1.33)	
	p-Value	27%	28%	68%	14%	62%	35%	20%	19%	
8 Months	DM Statistic	1.49	1.48	(0.49)	(1.18)	(0.64)	(1.07)	(1.34)	(1.37)	
	p-Value	14%	15%	63%	24%	52%	29%	19%	18%	

Panel C: Pairwise Diebold Mariano Tests by Product Line

		Asymmetric Optimal vs Official				Asymmetric Optimal vs Mean				Num
	γ	0.1	1	10	100	0.1	1	10	100	of Obs
Disti 1	DM Statistic	1.03	1.03	0.95	0.82	(0.51)	(0.97)	(1.33)	(1.34)	72
	p-Value	31%	30%	35%	41%	61%	34%	19%	19%	
Disti 2	DM Statistic	0.55	0.84	0.91	0.16	0.53	(0.51)	(1.30)	(1.36)	96
	p-Value	59%	40%	36%	87%	60%	61%	20%	18%	
Direct 1	DM Statistic	(1.57)	(1.62)	(1.51)	(1.45)	(0.52)	(0.94)	(1.28)	(1.32)	56
	p-Value	12%	11%	14%	15%	60%	35%	20%	19%	
Direct 2	DM Statistic	(1.61)	(1.45)	(1.33)	(1.42)	(0.37)	(0.82)	(1.25)	(1.32)	104
	p-Value	11%	15%	18%	16%	71%	41%	21%	19%	
Direct 3	DM Statistic	(1.02)	(1.03)	(1.14)	(1.25)	(0.20)	(0.58)	(1.04)	(1.13)	40
	p-Value	32%	31%	26%	22%	85%	56%	31%	26%	

asymmetric-loss-optimal forecast from the mechanism's distribution, denoted $\hat{Y}^{(\gamma)}_{i,t|t-h}$ for $\gamma \in \{0.1, 1, 10, 100\}$. We compare these asymmetric loss optimal forecasts with the official and the symmetric-loss-optimal mean forecasts using Diebold-Mariano Tests. The results in Table 7 confirm that the asymmetric-loss optimal forecasts continue to outperform the official forecast. Perhaps more interesting, though, is the extent to which an asymmetric-loss-optimal forecast outperforms the mean of the forecast distribution. This advantage in forecasting under asymmetric loss is unique to the information aggregation mechanism as implemented here and not available in any other prediction market mechanism.

8 Conclusion

In this paper, we introduce a new information aggregation mechanism in a novel field application forecasting practical business information needs. This implementation provides a testbed for evaluating the theory and refining the practice of information aggregation. We find such mechanisms are capable of capturing the true uncertainty of forward business indicators such as future sales and yield forecasts that are not only rational, but improve upon the forecast generated by internal processes. Further, by providing decision makers with a richer characterization of operational uncertainty, the mechanism can help them address problems with potentially asymmetric forecast loss and better control risks. As such, it is not surprising that Intel has extended its implementation beyond forecasting sales volumes but also to applications in evaluating research and development expenditures and tracking project management.

Field applications of information aggregation mechanisms also provide economists with valuable evidence for evaluating competing economic theories outside of tightly controlled laboratory settings. The IAM's successful implementation here further confirms the broad experimental evidence that such a mechanism can effectively aggregate information. A long process of testing and refining information aggregation mechanisms in the lab has borne real-world validation and value.

References

Mukhtar M. Ali. Probability and utility estimates for racetrack bettors. *Journal of Political Economy*, 85:803–815, 1977.

Beth Allen and James Jordan. The existence of rational expectation equilibrium: a retrospective. In M. Majumdar, editor, *Organization with Incomplete Information: Essays in Economic Analysis*. Cambridge University Press, 1998.

Kenneth Arrow, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest Nelson, George Neumann, Marco Ottaviani, Thomas Schelling, Robert J. Shiller, Vernon L. Smith, Erik Snowberg, Cass Sunstein, Paul Tetlock, Philip E. Tetlock, Hal R. Varian, Justin Wolfers, and Eric Zitzewitz. The promise of prediction markets. *Science*, 320:877–878, 2008.

Boris S. Axelrod, Ben J. Kulick, Charles R. Plott, and Kevin A. Roust. The design of improved parimutuel-type information aggregation mechanisms: Inaccuracies and the long-shot bias as disequilibrium phenomena. *Journal of Economic Behavior and Organization*, 69(2):170 – 181, 2009.

Joyce Berg, Robert Forsythe, Forrest Nelson, and Thomas Rietz. *Results from a Dozen Years of Election Futures Markets Research*, volume 1 of *Handbook of Experimental Economics Results*, chapter 80, pages 742–751. Elsevier, 2008.

Moez Bernnouri, Henner Gimpel, and Jacques Robert. Measuring the impact of information aggregation mechanisms: An experimental investigation. *Journal of Economic Behavior and Organization*, 78:302–318, 2011.

Adam Brandenburger and Ben Polak. When managers cover their posteriors: Making the decisions the market wants to see. *RAND Journal of Economics*, 27:523–541, 1996.

Kay-Yut Chen and Charles R. Plott. Information aggregation mechanisms: Concept, design, and implementation for a sales forecasting problem. *Mimeo*, 2002.

P. Chiappori, A. Gandhi, B. Salanie, and F. Salanie. Identifying preferences under risk from discrete choices. *The American Economic Review*, 99:356–362, 2009.

B. Cowgill, J. Wolfers, and E. Zitzewitz. Using prediction markets to track information flows: Evidence from Google. Working paper, 2009.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

Graham Elliott, Ivana Komunjer, and Allan Timmermann. Estimation and testing of forecast rationality under flexible loss. *Review of Economic Studies*, 72(4):1107–1125, 2005.

Graham Elliott, Ivana Komunjer, and Allan Timmermann. Biases in macroeconomic forecasts: Irrationality or asymmetric loss? *Journal of the European Economic Association*, 6(1):122–157, 2008.

Ray C Fair and Robert J Shiller. Comparing information in forecasts from econometric models. *American Economic Review*, 80(3):375–89, 1990.

A. Gandhi and R. Serrano-Padial. Are beliefs heterogeneous? the case of the longshot bias. working paper, 2012.

Clive W.J. Granger and Paul Newbold. *Forecasting Economic Time Series*. Academic Press, 1986.

Philip K Gray and Stephen F Gray. Testing market efficiency: Evidence from the nfl sports betting market. *The Journal of Finance*, 52:1725–1737, 1997.

David Harvey, Stephen Leybourne, and Paul Newbold. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291, 1997.

David I. Harvey, Stephen J. Leybourne, and Paul Newbold. Tests for forecast encompassing. *Journal of Business & Economic Statistics*, 16(2):254–259, 1998. ISSN 07350015.

Donald B Hausch and William T. Ziemba. *Handbook of Sports and Lottery Markets*. North-Holland, 2008a.

B. Jullien and B. Salanie. Estimating preferences under risk: The case of racetrack bettors. *Journal of Political Economy*, 108:503–530, 2000.

Katarina Kalovcova and Andreas Ortmann. Understanding the plott-wit-yang paradox. CERGE-EI Working Papers wp397, The Center for Economic Research and Graduate Education - Economic Institute, Prague, 2009.

Frederic Koessler, Charles Noussair, and Anthony Ziegelmeyer. Paramutuel betting under asymmetric information. *Journal of Mathematical Economics*, 44:733–744, 2008.

Frederic Koessler, Charles Noussair, and Anthony Ziegelmeyer. Information aggregation and belief elicitation in experimental parimutuel betting markets. *Journal of Economic Behavior and Organization*, 83:195–208, 2012.

C. Manski. Interpreting the predictions of prediction markets. *Economics Letters*, 91:425–429, 2006.

P. Milgrom and N. Stokey. Trade, information and common knowledge. *Journal of Economic Theory*, 26:17–27, 1982.

Jacob A. Mincer and Victor Zarnowitz. The evaluation of economic forecasts. In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, NBER Chapters, pages 1–46. National Bureau of Economic Research, Inc, 1969.

Marco Ottaviani and Peter Norman Sørensen. Noise, information, and the favorite-longshot bias in parimutuel predictions. *American Economic Journal: Microeconomics*, 2(1):58–85, 2010.

Andrew J. Patton and Allan Timmermann. Forecast rationality tests based on multi-horizon bounds. *Journal of Business & Economic Statistics*, 30:1–17, 2012.

C. R. Plott and K.A. Roust. The design of information aggregation mechanisms: A two-stage parimutuel market to avoid mirages bubbles. *Mimeo*, 2009.

Charles R. Plott. Markets as information gathering tools. *Southern Economic Journal*, 67 (1):1–15, 2000. ISSN 00384038.

Charles R. Plott and Shyam Sunder. Efficiency of experimental security markets with insider information: An application of rational-expectations models. *Journal of Political Economy*, 90(4):663–698, 1982. ISSN 00223808.

Charles R. Plott and Shyam Sunder. Rational expectations and the aggregation of diverse information in laboratory security markets. *Econometrica*, 56(5):1085–1118, 1988. ISSN 00129682.

Charles R. Plott, Jorgen Wit, and Winston C. Yang. Parimutuel betting markets as information aggregation devices: experimental results. *Economic Theory*, 22(2):311–351, 2003.

Erik Snowberg and Justin Wolfers. Explaining the favorite-long shot bias: Is it risk-love or misperceptions? *Journal of Political Economy*, 118(4):723–746, 2010.

Allan Timmermann. Forecast combinations. In C.W.J. Granger G. Elliott and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 1 of *Handbook of Economic Forecasting*, chapter 4, pages 135 – 196. Elsevier, 2006.

Kenneth D. West. Forecast evaluation. In C.W.J. Granger G. Elliott and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 1 of *Handbook of Economic Forecasting*, chapter 3, pages 99 – 134. Elsevier, 2006.

J. Wolfers and E. Zitzewitz. Prediction markets in theory and practice. IZA Discussion Paper #1991, 2006.

Linda M Woodland and Bill M Woodland. Market efficiency and the favorite-longshot bias: The baseball betting market. *Journal of Finance*, 49:269–279, 1994.

Not for Publication: Web Appendix

A1. Monotonicity Tests across Forecast Horizons

One interesting feature apparent in the forecasts, after correcting for scale distortion, they tend to be conservative on average despite the overall positive bias we observe in the summary statistics from table 1. Though this feature is not statistically significant, it is widespread across forecast horizons and product lines, indicating forecasting loss may have some asymmetric properties. Another interesting feature from these results finds the location distortion decreases in the horizon, as illustrated in Figure 3.

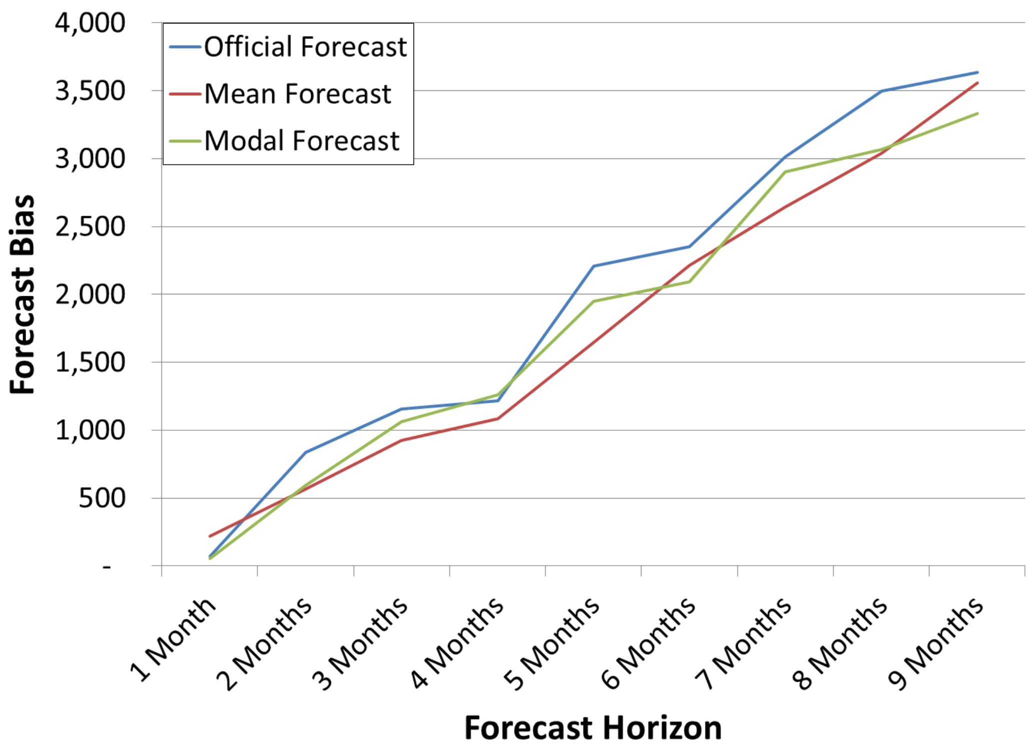


Figure 3: Mincer-Zarnowitz Location Distortion by Forecast Horizon
This figure summarizes the intercept coefficients estimated from the Mincer-Zarnowitz regressions in table 2.

Beyond the pooled MZ test, rational forecasting across multiple horizons imposes a number of additional testable restrictions on the second moments of forecasts, forecast revisions, and forecast errors. Several of these restrictions take the form of monotonicity restrictions in the forecast horizon. For example, the mean square forecast error should be increasing in the horizon h as there is greater uncertainty with longer horizons. Similarly, the covari-

Table 8: Forecast Rationality Tests of Multiple Horizon Restrictions

This table presents the p-Values for hypothesis tests that evaluate the null hypothesis that the population restriction in the row holds under the forecast identified in each column.

	Official Forecast		Mean Forecast		Mode Forecast	
	9 Month vs 1 Month	Overall	9 Month vs 1 Month	Overall	9 Month vs 1 Month	Overall
Increasing MSFE	93%	100%	93%	100%	95%	100%
Decreasing $\mathrm{Cov}(Y, Y_h)$	64%	35%	49%	34%	27%	34%
Decreasing MSF	53%	30%	32%	19%	16%	15%
Increasing MSFR	66%	70%	69%	65%	75%	60%
Bonferroni Combined Test	60%	26%	35%	16%	12%	12%
Number of Obs	46					

ance between forecasted sales and actual sales should be decreasing in the forecast horizon as increased uncertainty allows a larger wedge to form between the forecasted and realized outcome.

Patton and Timmermann (2012) (PT) provide a detailed analysis of these restrictions and propose a suite of tests to evaluate their validity. Indeed, they show that rational forecasting imposes restrictions that almost every pairwise relationship of outcomes, forecasts, forecast errors, and forecast revisions be either monotonic in forecast horizon or bounded by fixed values. We focus on four of these monotonicity restrictions, namely that the Mean Square Forecast Error (MSFE) should increase for longer horizons. Also, if long horizon forecasts are less accurate, we should see a decrease in the covariance between the forecast and the outcome. Similarly, larger forecast revisions should be required at these longer horizons and the overall magnitude of the forecasts should diminish.

We follow PT in constructing sample analogs to the population moments and calculate p-Values using the bootstrap test proposed by Patton and Timmermann (2008) for testing monotonic conditional expected returns across a set of sorted portfolios. We report the p-Values for the four relevant restrictions and Bonferroni-corrected p-Values for a combined test of all restrictions in Table 8.

These tests do not reject the null hypothesis of forecast monotonicity, however it's difficult to evaluate the extent to which this failure is due to insufficient power rather than well-ordered forecast behavior. Since the overall tests seek to evaluate eight restrictions using only 46 observations, absent severe forecast inefficiencies, we'd be surprised to find high levels of significance.

A2. Harvey, Leybourne, and Newbold Encompassing Tests

To test the robustness of the results from the encompassing tests in section 6.2, we implement a set of pairwise encompassing tests as proposed by Harvey, Leybourne, and Newbold (1998, henceforth HLN). The HLN encompassing test modifies the DM test of forecast comparison to test if the difference in the errors for the reference and alternative forecasts is correlated with the errors in the reference forecast. If this correlation is non-zero, then a forecast user could exploit that correlation to reduce their loss by combining the two forecasts, though when the HLN statistic is negative, the optimal forecast requires forming a "hedge" that weights the reference forecast positively while negatively weighting the alternative.

As in developing the DM test statistic, we start by defining a pairwise measure of covariance between forecast error and the difference in forecast error at each period:

$$\eta_{i,t,h}^{(j,k)} = \left(e_{i,t|t-h}^{(j)} - e_{i,t|t-h}^{(k)}\right) e_{i,t|t-h}^{(j)} \tag{12}$$

Then, in parallel to the DM test statistic, the HLN statistic is the t-statistic for testing whether the average $\eta_{i,t,h}^{(j,k)}$ differs from zero. Of course, the studentization requires an estimate of the variance robust to autocorrelation in errors and allowing for clusters by period. As with the DM test statistic, the finite-sample properties of the HLN test statistic are better modeled using the Student's t-Distribution.

In the table 9, we present the HLN Test statistics and p-Values for each pairwise combination of tests. The first two columns test the null hypothesis that the Mean Forecast does not encompass the Official Forecast and Mode Forecast, respectively. In these results, it is apparent that the Official Forecast typically does not encompass the Mean or Mode Forecast. Similarly, the Mean and Mode forecasts fail to encompass by the official forecast, indicating that the mechanism is not simply restating the official forecast but incorporates additional information. As should be expected, we cannot reject the hypothesis that the Mean and Mode forecasts are encompassed by each other. This overall pattern of relationships between forecasts is fairly robust to forecast horizon and product line subsamples.

A3. Experimental Instructions

Table 9: Encompassing Tests across Forecast Mechanisms

This table presents Harvey, Leyborne, and Newbold (1998) tests comparing the point forecasts from the official forecast and the mean and mode parimutuel mechanism forecast. Panel A uses the full sample of all forecasts and horizons, with Panels B and C reporting results for horizon and product subsamples. The HLN Statistic and associated p-Value use standard errors robust to autocorrelation up to the maximum horizon included in the sample, clustered by period and product.

Panel A: Full Sample Harvey, Leybourne, Newbold Tests

Reference	Official		Mean		Mode		Num of
Alternative	Mean	Mode	Official	Mode	Official	Mean	Obs
HLN Statistic	7.20***	7.19***	3.96***	0.87	4.06***	1.57	414
p-Value	0%	0%	0%	38%	0%	12%	

Panel B: Harvey, Leybourne, Newbold Tests by Forecast Horizon

Reference	Official		Mean		Mode		
Alternative	Mean	Mode	Official	Mode	Official	Mean	
h = 1	3.50***	3.54***	1.23	1.11	1.29	(0.42)	46.00***
	0%	0%	23%	27%	20%	68%	
h = 2	3.32***	3.29***	1.68*	1.39	1.45	0.88	
	0%	0%	10%	17%	15%	38%	
h = 3	3.16***	3.15***	1.17	1.67	1.37	(0.52)	
	0%	0%	25%	10%	18%	60%	
h = 4	3.37***	3.33***	0.50	0.16	0.52	0.39	
	0%	0%	62%	87%	60%	69%	
h = 5	3.05***	3.05***	1.19	1.50	0.89	(0.64)	
	0%	0%	24%	14%	38%	52%	
h = 6	1.95*	1.80*	1.62	(1.43)	2.00*	1.98*	
	6%	8%	11%	16%	5%	5%	
h = 7	0.91	1.02	1.95*	0.45	1.98*	0.63	
	37%	31%	6%	66%	5%	53%	
h = 8	0.63	0.50	2.48 **	(0.05)	2.48 **	0.82	
	53%	62%	2%	96%	2%	41%	

Panel C: Harvey, Leybourne, Newbold Tests by Product Line

Reference	Official		Mean		Mode		
Alternative	Mean	Mode	Official	Mode	Official	Mean	
Disti 1	2.23 **	1.74*	3.92***	(0.80)	3.86***	1.67*	72
p-Value	3%	9%	0%	43%	0%	10%	
Disti 2	2.67***	2.63***	3.93***	1.04	3.88***	0.81	96
p-Value	1%	1%	0%	30%	0%	42%	
Direct 1	3.46***	3.45***	1.49	(0.70)	1.84*	2.01 **	56
p-Value	0%	0%	14%	48%	7%	5%	
Direct 2	5.76***	5.69***	1.74*	0.26	2.02 **	1.34	104
p-Value	0%	0%	8%	79%	5%	18%	
Direct 3	4.34***	4.59***	1.96*	2.01*	1.46	(1.10)	40
p-Value	0%	0%	6%	5%	15%	28%	

*, **, *** indicates significance at the 10%, 5%, and 1% confidence levels, respectively.

Forecasting Instructions – (date and time)

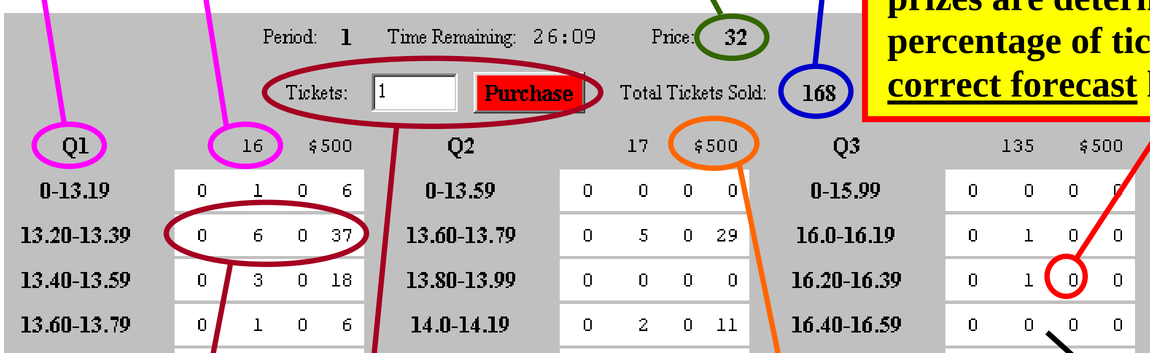
Forecasting (variable to be forecast)

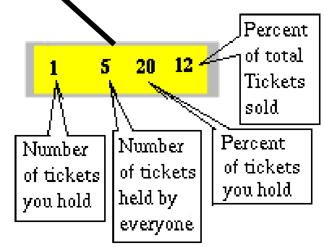
Each column lists the set of forecasts for one quarter and total tickets sold

The price of a ticket will start to increase 15 minutes into the session

Total tickets sold to all participants for all quarters

Your chances of winning prizes are determined by the percentage of tickets <u>in the correct forecast</u> held by you





To purchase a ticket:
1. Click the white box of the range you choose
2. Enter the number of tickets
3. Click Purchase

Your unspent cash used to purchase tickets – compare to ticket price – separate budget for each quarter/column

Strategy:
1. You start with 500 units of house money for each quarter. <u>Spend it all</u> – but not on one forecast range unless you are certain.
2. Watch what others are doing. <u>The objective is to win money, not simply to record your beliefs.</u>
3. Prices will start at 5 units/ticket and not change for the first 15 minutes. Then they will go up by one unit per minute for 45 minutes. Do not wait too long to buy.

<u>Practice</u>:
<u>http://location</u> and time
<u>Real Deal</u>:
Time and location

Procedure

Step 1: Register

Register yourself in the system database. If you are not in the database the system will force you to register when you try to log into the Real Deal.

Go to (at any time including now) http://xxxx.caltech.edu/xxx
Select "Sign up as a new user". Choose an ID, a password, and enter a number into the "SS Number" field. We are not using real social security numbers – just pick a number with 9 digits that you can remember (or write down). Part of a phone number might be a good idea.

Everyone should enter the following information. It will not be used for anything but is required in the stock application we are using.

University = "Company A" and Class = "Company A"
Street = "123 Main Street" City = "Anytown"
State = "CA" Zip = "12345" Country = "USA"
Enter your real e-mail address and phone number. (Enter area code "123" and then your real seven digit Intel phone number.)

Step 2: Practice

Go to the practice page http://xxxx.caltech.edu/Sales-practice/ prior to the Real Deal to become familiar with the forecasting application. Buy tickets for a few different forecasts and observe how the application responds.

Step 3: Get your secure ID

On the day of the Real Deal, ideally a few minutes before the start time, go to the Real Deal location, http://xxxxcaltech.edu/BusinessUnitYearQ#Date/. It will ask you for the user name and password that you used in Step 1. It will then give you your secure ID, which disguises your identity. Click the "Login" button to enter the Real Deal. You will not be able to use the application until the session begins.

Step 4: Participate in the Real Deal

The session will be held on November 7 at 4:00 PM Pacific Time. Be on time – a few minutes early would be wise. The trial will start exactly on time, allowing for clock differences, and move very quickly. It will likely be over in 30 minutes even though it will remain open for an hour.

Panics or problems: e-mail or call Mister X at ###-###-####. He will be working with Caltech to manage the trial and solve any problems.

We will put general announcements (if needed) on the Real Deal screens.

Determining Winners

Four prizes will be awarded for each of the three quarters forecast during the trial – see details below. We will know which forecast is correct once actual Q4 2006 and Q1, Q2 2007 Business Unit Billings are available. Prizes for each quarter will be awarded after the close of that quarter. All tickets in the correct forecast are considered winning tickets and will be entered into a drawing for prizes. After each prize drawing the winning ticket will be put back in the hopper, so each ticket may win more than one prize.

Q4 2006	Q1 2007	Q2 2007
Drawing 1: $100	Drawing 1: $100	Drawing 1: $100
Drawing 2: $100	Drawing 2: $100	Drawing 2: $100
Drawing 3: $50	Drawing 3: $50	Drawing 3: $50
Drawing 4: $50	Drawing 4: $50	Drawing 4: $50

These prizes will be distributed as an employee recognition award in the near term. Alternative payment methods may be developed in the long term.

Privacy

Participants will remain completely anonymous except to the research team at Caltech and to Mister X, the research manager at Company A. No one else participating in the trial will know for certain who is participating, so they certainly will not know which forecasts you choose. The final forecast generated by all participants will be published, but your personal forecast will be held in confidence by the research team. We will award prizes to the winners, but even the winners will not be announced.

We expect that participants will not share information with one another before, during or after the trial. Past research has shown that the best results are achieved when participants do not share information.