Comment

Christopher D. Carroll, Johns Hopkins University and NBER

This paper opens up what promises to be a whole new approach to macroeconomic research. Market-based forecasts of macroeconomic variables provide a promising way to neatly sidestep the intractable, insoluble, and semi-theological debates about how expectations are formed that have plagued macroeconomics since Keynes first speculated that "animal spirits" were a driving force in business cycles.

So you might say I'm a fan.

In fact, the first part of my discussion will argue that the results of the paper are even more important than one might conclude from the authors' own analysis, because they focus on the (microscopic) differences between survey-based forecasts and market-based forecasts, rather than on the impressive similarities between them. The brief latter part of the discussion raises some reasons for caution about the institutional design and operation of these markets.

1. Comparing Survey and Auction Based Expectations

A substantial part of the paper (Tables 1–3) compares expectations as revealed by the auction market to the mean forecasts of a survey of professional forecasters. An incautious reader might get the impression that these results suggest the market-based expectations are notably better than those of the survey. In fact, I think the opposite interpretation is the right one: When used to measure the same thing, survey-based expectations are, for analytical purposes, indistinguishable from market-based expectations.

Consider, for example, the non-farm payrolls data, which are for most purposes the most important single U.S. data release.[1] The authors present the following comparative statistics about the two. (These are taken from their Table 1).

Table 1
Prediction errors from auction and survey (non-farm payrolls)

	Mean absolute error (AbsErr)	Root mean squared error (RMSE)	Correlation with actual outcome
Auction	0.723	0.907	0.700
Survey	0.743	0.929	0.677

The table speaks for itself.

The authors emphasize the results for their other data series, which could be described as providing a smidgen of evidence that the market forecasts are more accurate than the survey forecasts. I will shortly express some quibbles with this interpretation. But before doing so, I would like to point out that even under the authors' interpretation, the superiority of the auction forecast is generally small.

This is important because the macroeconomic derivatives markets have been operating only for a short time. Since, according to the NBER Business Cycle Dating Committee, the average postwar business cycle in the U.S. has had a duration of about eight years, the usefulness of these data for macroeconomic analysis will arguably be modest for at least a decade. If instead we draw the conclusion that the macroeconomic derivatives markets have definitively revealed the impressive qualities of survey-based expectations, the scope of the paper's usefulness is vastly expanded, since various kinds of survey-based expectations have been collected for a very long time (for example, the Survey of Professional Forecasters has been conducted since 1968).

1.1 Quibbles

As the authors note, the auctions they analyze do not provide any real opportunity for hedging macroeconomic risks in the sense Shiller (1993) originally proposed because they are generally conducted only a few hours (or at most a few days) before the data are released.

This timing, however, means that participants in the auctions have more recent information than survey participants, whose views are collected every Friday. In the case of a data series released on a Thursday, the auction participants' information set could incorporate nearly a week's worth of extra knowledge about the state of the economy.

This problem is particularly serious for initial claims for unemployment insurance, since this is a weekly series released on Thursday

mornings. Indeed, it is remarkable that the almost week-old surveys do almost as well as the previous-day auctions in forecasting this weekly series.

An alternative way of analyzing the authors' data (and one that is fairer to the forecasters) would be to hypothesize that both forecasters' and auction participants' views are rational; in that case, Hall (1978) taught us that the auction results should equal the survey results plus a random expectational error that reflects the forecasters' extra information:

$$A_t = S_{t-1} + \varepsilon_t,$$ (1)

which can be tested by estimating a regression

$$A_t = z_0 + z_1 S_{t-1}$$ (2)

and testing $z_0 = 0$ and $z_1 = 1$.

To test this proposition as an overall characterization of the authors' data, it is necessary to put the various statistics on a common footing in the sense of having comparable means and measures of variability. I did so by subtracting, for each series, the mean realized value over the sample period, and dividing by the gap between the maximum and minimum realized sample values.[2]

Results are plotted in Figure 1. As the figure illustrates, there is a very strong association between the survey and the auction predictions.

The point is illustrated statistically by Table 2, which reports the results of a regression like the one contemplated in equation (2). The hypotheses that $z_0 = 0$ and $z_1 = 1$ cannot be rejected at standard significance levels, and the $\bar{R}^2$ for the regression is over 90 percent. When the sample is restricted to the crucial non-farm payrolls data, similar results obtain.

One way of testing whether the more up-to-date information held by auction market participants could plausibly explain a modest superiority in their forecasts is to see whether auctions that are held closer to the date of the data release produce forecasts that are more accurate. Unfortunately, the authors' dataset contains only a few auctions that were held earlier than the day on which a data series was released. Most of these were for the ISM data. Table 3 calculates the size of the absolute error for the 21 auctions that were held on the morning of the data release, the four auctions that were held one day before, and the three auctions that were held three days before. (There seem to be no examples of auctions conducted two days before the release). The mean absolute error is notably larger for the auctions conducted rela-
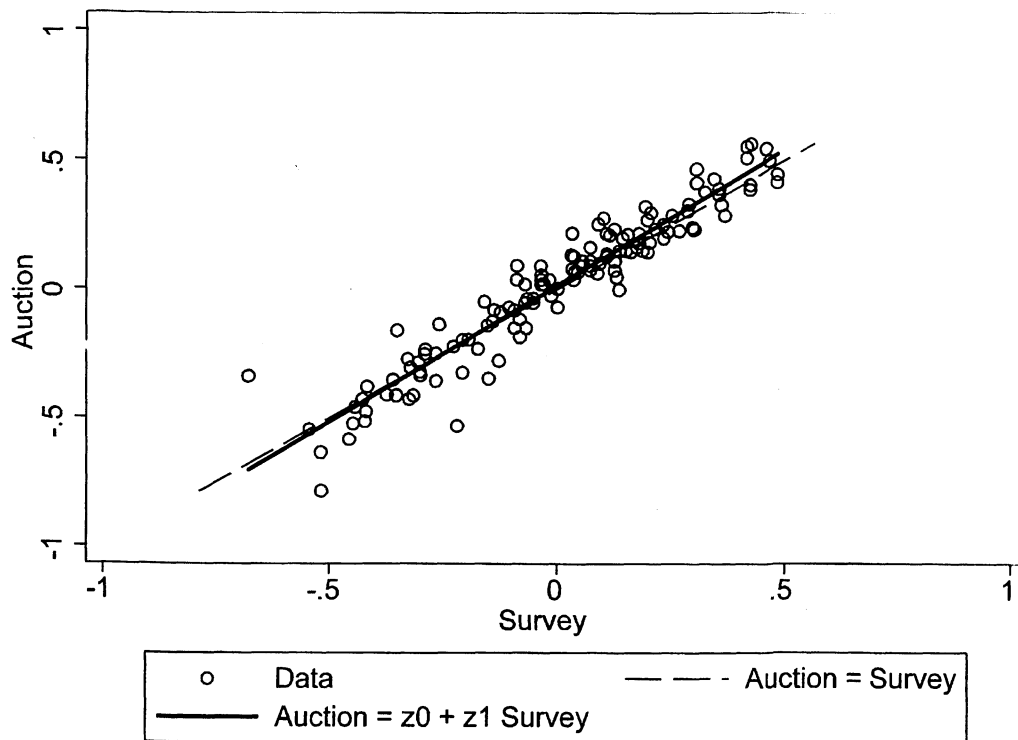
Figure 1
Survey expectations versus auction expectations

Table 2
Regression of auction on survey expectations

Data series	Auction = z0 + z1 Survey		
	z0	z1	$\bar{R}^2$
All	0.013	1.055	0.91
	(0.007)	(0.039)	
Payrolls	0.001	1.096	0.95
	(0.014)	(0.052)	

Robust standard errors in parentheses.

Table 3
Absolute error for different ISM auction horizons

Days between auction and data release	Number of auctions	Mean absolute error
0	21	0.48
1	4	0.57
3	3	0.56

tively earlier, as would be true if significant news generally arrives in the period leading up to the release (though separate tests (not shown) indicate that these differences are not statistically significant).

The authors emphasize the results of a final horse race (in Table 2) between the two series. They show (convincingly) that financial market reactions to the actual data release are stronger when the "surprise" is measured as the deviation from the auction forecast than when it is measured as the deviation from the survey forecast, at least for the payrolls data.

Again a possible explanation is the later date of the auction than the survey. Another possibility that the authors suggest is that the participants in the auctions are precisely the same people whose financial transactions, post-release, will determine the market reaction. If this is true, it would be puzzling if their opinions did *not* have more influence on financial market outcomes than the opinions of bystanders like the economists participating in the surveys.

None of this is meant to dispute the proposition that the auction based forecasts are a superior source of information, when both auction and survey data exist. As the authors show, the auction data paint a much richer picture of expectations than is available from the surveys, particularly with respect to the probability distribution over possible outcomes, which can be condensed (as the authors show) in any of several ways to measure uncertainty. In 30 years there may be no reason to use survey data at all because a sufficient amount of auction data will be available. But for the time being, the authors' results provide compelling evidence that surveys capture an enormous amount of useful information.

This richness is used in section 4 of the paper to examine a question that heretofore has been a matter of speculation: whether disagreement among survey participants can be interpreted as a measure of uncertainty.

On the whole their conclusion is that such an interpretation is problematic. Table 4 reproduces the key results from their analysis of this question, in which they regress measures of uncertainty on measures of disagreement. The absolute magnitudes of the coefficients are not meaningful, because there is no obvious mapping between the cross-forecaster standard deviation of forecasts of the mean value of the release, and the standard deviation of the released data itself. The right questions are the degree of statistical significance of the relationship

Table 4
Uncertainty versus disagreement

Series	Uncertainty = $\alpha + \beta$ Disagreement	
	β	$\bar{R}^2$
Payrolls	0.66**	0.11
	(0.29)	
Retail sales	0.44**	0.20
	(0.16)	
Initial claims	0.27***	0.17
	(0.07)	
ISM	−0.03	−0.03
	(0.12)	

between uncertainty and disagreement, and the total proportion of uncertainty that can be measured by disagreement. Except for the ISM series, the authors find a highly statistically significant relationship between disagreement and uncertainty.

They tend to emphasize, however, the finding that the $\bar{R}^2$ is well below one in all cases. But there is clearly sampling error in the survey of forecasters; how to think about this is not entirely obvious, since there are forecasters who exist but are not in the survey and the survey participants vary over time. By itself this would be enough to prevent an $\bar{R}^2$ equal to one even if the authors' measures of uncertainty were perfect.

My own sense is that the more important question is whether disagreement can be interpreted as a statistically reliable indicator of the degree of uncertainty, rather than a direct measure. One way to make the question concrete is to ask whether the regression the authors report can be thought of as the first stage of a two-stage least squares regression of uncertainty on disagreement. One could then use the prediction of the estimated equation as a contemporaneous measure of appropriately calibrated uncertainty. Judged in this way, the $\bar{R}^2$'s for the first stage regressions and the high statistical significance of the coefficients are plenty good enough to interpret the prediction of the model as an (instrumented) measure of uncertainty. (Of course, careful econometrics would have to make sure that this cross-section disagreement is not perfectly correlated with some other macro variable (like the inflation rate).)

2. Caveats about Macro Markets

Despite their many attractive properties, it is worth worrying a little bit (at this early stage) about the longer term consequences of the creation of macro markets, especially for the data collection process.

I have the fullest faith in the integrity and objectivity of the staff at the agencies that produce economic data. But there can be no doubt that the creation of macro markets will increase both the pressure on the staff and the ease with which an unscrupulous employee could exploit inside information. Data security procedures need not only to be objectively rigorous but also to be transparently seen to be rigorous. Possibly there should be a systematic ongoing program (by the Securities and Exchange Commission?) to monitor trading in macro markets for any signs of insider trading.

Another concern is that if macro markets become sufficiently popular (and lucrative), the economic agencies may have a problem of retaining senior staff. If senior officials were regularly lured away from their posts by the offer of salaries many times higher than the government can provide, it might be difficult to preserve the institutional memory and expertise necessary for guaranteeing the consistency and high quality of U.S. statistics. Probably the only appropriate measure that could be taken to prevent this (in addition to paying appropriately high salaries to the senior staff) would be to impose strict ethics rules that require a substantial waiting period (say, five years) between the time of departure from a statistical agency and any employment that exploits that expertise in the context of macro markets.

Finally, and perhaps most significantly, the existence of macro markets could influence the data collection procedures themselves. Although the currently existing auction markets probably do not pose much risk in this dimension, when markets are created for longer-term forecasts (as they inevitably will be), the holders of those auction contracts will have the incentive to become lobbying groups for or against changes in the methods of data collection. Imagine, for example, that macro markets had existed at the time of the Boskin Commission on reform of the CPI in the mid-1990s, or the redefinition of the unemployment rate in the early 1990s. If each decision a commission announces results in immediate capital gains or losses of billions of dollars for holders of contingent securities, there will be extraordinary incentives to subvert the objectivity of the decision makers. Good institutional design could

certainly circumvent these pressures, but if data collection procedures are perceived to be able to be influenced by the appointment of ad hoc committees nominated by politicians there is reason to worry.

This risk could perhaps be alleviated if the agencies that produce the data were to create standing committees of scientific advisors associated with each of the major statistical releases for which macro markets exist or are in contemplation. For example, a panel of distinguished labor economists might be recruited to monitor proposed changes to the non-farm payrolls survey. These committees might borrow the model of the NBER Business Cycle Dating Committee: Meetings only when warranted by some event, but a committee that is always well defined. This would provide some transparent insulation against the political forces that might otherwise mobilize to have commissions appointed whose members would be picked to reach preordained conclusions.

It is important to resolve these issues early, because the whole superstructure of macro markets will be undermined if the integrity of the data collection process comes into question. But if addressed early, these problems should not be serious.

3.   Conclusions

All quibbles aside, this paper, and the macro markets that it is the first to explore, represent a tremendous innovation in macroeconomic analysis. I look forward with great anticipation to the literature that will undoubtedly flow from them.

Notes

1. Like the authors, Fleming and Remolona 1997 find that this data release moves the bond market more than any other, and more recently Faust et. al. 2003 have found that this data release moves exchange rates even more than monetary policy surprises.

2. Results were similar when the data were scaled, following the authors, by the presample standard error; the resulting figure is slightly more legible using my scaling method.

References

Faust, Jon, John H. Rogers, Eric Swanson, and Jonathan H. Wright. 2003. "Identifying the Effects of Monetary Policy Shocks on Exchange Rates Using High Frequency Data." *Journal of the European Economic Association* 1: 1031–1057.

Fleming, Michael, and Eli Remolona. 1997. "What Moves the Bond Market?" *Federal Reserve Bank of New York Economic Policy Review* 3: 31–50.

Hall, Robert E. 1978. "Stochastic Implications of the Life-Cycle/Permanent Income Hypothesis: Theory and Evidence." *Journal of Political Economy* 96: 971–987.

Shiller, Robert J. 1993. *Macro Markets: Creating Institutions for Managing Society's Largest Economic Risks.* Oxford: Oxford University Press.

Comment

Adam Szeidl, University of California, Berkeley

1. Introduction

This is an interesting and informative paper that explores pricing behavior in a new market for macroeconomic derivatives. Asset markets where risk associated with future macroeconomic events can be traded are a recent financial innovation. These markets may allow more efficient sharing of macro risks and increase economic welfare. To assess their potential, it is important to understand how well existing economic derivatives markets function. Analyzing data from one such market where claims on macroeconomic indicators including non-farm payrolls are traded, this paper argues that (1) Expectations derived from market prices are more accurate than survey-based forecasts and less subject to behavioral biases; (2) The market predicts the probability distribution of outcomes remarkably well; (3) Risk aversion plays at most a small role in determining prices in this market.

I begin by discussing potential theoretical foundations for the empirical findings. Then I briefly discuss features of the market mechanism, and finally turn to the role of risk aversion. My comments suggest additional empirical tests that can sharpen our understanding of how markets for economic derivatives function.

2. Theory

Perhaps surprisingly, it is not easy to come up with plausible microfoundations for findings (1) and (2). Why are prices accurate predictors of outcomes? And why are prices more accurate than survey-based forecasts, when in many economic models, prices are functions of the beliefs that forecasts measure? To answer these questions, I begin by exploring the mechanism through which markets may aggregate infor-

mation. A large theoretical literature (e.g., Grossman 1976 or more recently Reny and Perry 2003) argues that markets correctly aggregate heterogeneous information in the presence of common prior beliefs. In practice, however, the common prior assumption appears to be at odds with often-observed disagreement in survey forecasts among professional forecasters, because different individuals with common priors cannot agree to disagree (Aumann 1976). A plausible alternative in this context is to assume that disagreement is due to heterogeneous prior beliefs.

However, with heterogeneous beliefs, as argued for example by Manski (2004), it is not a-priori clear that predictive markets should correctly aggregate information. To see the logic, note that in principle, a wealthy individual with incorrect beliefs may be able to push prices away from fundamental values by the sheer size of her investment. More formally, Wolfers and Zitzewitz (2005) show that with risk-averse investors and a competitive market, the price will equal the wealth-weighted average belief in the population. This result confirms that market prices can depart from true expectations if the distribution of beliefs is correlated with wealth. On the other hand, in this model, accurate market prices obtain if the average belief in the population correctly predicts outcomes. This suggests that the reason why predictive markets function so well is that the average belief of investors is correct.

To test this proposition, one can look for alternative empirical measures of beliefs. A natural candidate, used for example by Mankiw, Reis, and Wolfers (2003), is survey-based forecasts. If one accepts that such surveys are a good measure of beliefs, then the Wolfers-Zitzewitz model predicts that surveys will forecast outcomes at least as well as market prices. However, this prediction contradicts finding (1) of this paper. How can prices be more accurate than surveys, when surveys are a direct measure of investors' beliefs?

To resolve this contradiction, one has to relax one of the assumptions of the previous argument. It must be that either (a) prices are not more accurate than survey-based forecasts; or (b) surveys do not reflect true beliefs; or (c) prices are accurate not because they reflect average beliefs, but for some different reason. Distinguishing between these alternatives would be useful to better understand the workings of predictive markets.

Let us address each possibility in turn. Case (a) suggests that finding (1) in the paper is due to other differences between the survey and market data. Timing is one such difference: while the predictive market

meets on the morning of the data release, the survey is collected up to a week earlier. Given such differences in timing, information that becomes available after the survey is collected may be reflected in the market price. This explanation suggests that surveys are good measures of expectations. From a practical perspective, this would be useful, because survey data is more widely available than data from predictive markets. Using the data of the current paper, this explanation can be tested by comparing the differential accuracy between surveys and forecasts depending on the difference in timing. When this explanation is correct, surveys that take place later should be closer in accuracy to market prices.

Case (b) may hold for example if survey respondents have little to lose from making incorrect predictions, while market participants have money at stake. In this case, earlier work where beliefs are measured using survey based forecasts is potentially misleading. While there is little doubt that predictions do improve when the stakes are higher, the question is quantitative. How much does precision increase when the stakes go up? A preliminary empirical approach to explore this question is to compare the accuracy of predictions across markets with different stakes, as measured perhaps by total investment in short and long positions. In markets with higher total investment, we should find that prices are better predictors of outcomes.

In my view, case (c) is the least likely. If prices do not reflect average beliefs, then we are back to the original puzzle: Why do prices in predictive markets forecast outcomes so accurately?

To summarize, the most plausible theory raises the question of whether finding (1) is caused by the different nature of surveys versus markets or their differential timing, and suggests additional empirical tests to help sort out whether markets are just as accurate as surveys or more accurate because the stakes are small for survey participants.

3.   The Pari-Mutuel Mechanism

Understanding the logic of information aggregation in predictive markets is further complicated by the fact that the market mechanism is not competitive. The market is a modified version of the pari-mutuel mechanism often used in horse race betting. Eisenberg and Gale (1959) explore Nash equilibrium in a simple version of the basic pari-mutuel model. They establish existence and uniqueness of equilibrium; how-

ever, the equilibrium they find need not involve prices that correctly predict outcomes. To quote the last sentence in their paper: "In the case of two bettors with equal budgets if the first bettor's subjective probability distribution on two horses is ((1/2),(1/2)) then the equilibrium probabilities will be ((1/2),(1/2)) regardless of the subjective probabilities of the second bettor, as the reader will easily verify." Therefore, in the special case discussed in the quote, the price will be independent of the beliefs of the second bettor. This example suggests that exploring the actual market mechanism in more detail can lead to useful insights about the logic of information aggregation.

4. Risk Aversion

My final topic is the role of risk aversion. Using a simple model with power utility investors, the paper shows that for reasonable coefficients of relative risk aversion the risk premium of holding economic derivatives should be very small. Based on this argument, the authors conclude that risk is unlikely to affect asset prices in predictive markets.

One problem with this logic is that the same calibration argument, if applied to the aggregate stock market, would imply that risk plays at most a minor role in determining expected stock returns, and that the equity risk premium should be very small. As it is well known, this implication of the model is robustly contradicted in the data (e.g., Mehra and Prescott 1985). This equity premium puzzle suggests that the standard power utility model should not be used to assess the effect of risk in influencing asset prices. An alternative approach to gauge the impact of risk on prices is to note that for most investors, investing in predictive markets is likely to be a relatively small risk. There are studies suggesting that decision making in the presence of small risk is well-described by loss-aversion preferences that have a kink at the status quo level of wealth (see for example, Thaler, Tversky, Kahneman, and Schwartz 1997). Calibrating a model with such loss-averse investors would be an empirically more plausible way to assess the role of risk in affecting predictive market prices.

To conclude, this is an interesting paper that documents useful facts about the functioning of economic derivatives' markets. I hope that my discussion helps in suggesting additional empirical tests to sharpen our understanding of the mechanism through which these markets aggregate information.

References

Aumann, Robert. 1976. "Agreeing to Disagree." *The Annals of Statistics* 4: 1236-1239.

Eisenberg, Edmund, and David Gale. 1959. "Consensus of Subjective Probabilities: The Pari-mutuel Method." *Annals of Mathematical Statistics* 30: 165–168.

Grossman, Sanford. 1976. "On the Efficiency of Competitive Stock Markets Where Trades Have Diverse Information." *Journal of Finance* 31: 573–585.

Mankiew, Gregory N., Ricardo Reis, and Justin Wolfers. 2003. "Disagreement about Inflation Expectations." *NBER Macroeconomics Annual*.

Manksi, Charles F. 2004. "Interpreting the Predictions of Prediction Markets." Working Paper no. 10359. Cambridge, MA: National Bureau of Economic Research.

Mehra, Rajnish, and Edward C. Prescott. 1985. "The Equity Premium: A Puzzle." *Journal of Monetary Economics* 15: 145–161.

Reny, Philip J., and Perry, Motty. 2003. "Toward a Strategic Foundation for Rational Expectations Equilibrium." Working Paper. University of Chicago and Hebrew University.

Thaler, Richard H., Amos Tversky, Daniel Kahneman, and Alan Schwartz. 1997. "The Effect of Myopia and Loss Aversion on Risk Taking: An Experimental Test." *Quarterly Journal of Economics* 112: 647–661.

Wolfers, Justin, and Eric Zitzewitz. 2005. "Interpreting Prediction Market Prices as Probabilities." Working Paper. University of Pennsylvania and Stanford University.