INTERPRETING AND COMBINING HETEROGENEOUS SURVEY FORECASTS

Charles F. Manski
Department of Economics and Institute for Policy Research
Northwestern University

Forthcoming: *Oxford Handbook on Economic Forecasting*, M. Clements and D. Hendry (editors)

This research was supported in part by National Science Foundation grant SES-0911181. I am grateful for comments from Mike Clements.

1. Introduction

This handbook chapter concerns forecasts reported in surveys. I exposit several logical issues that arise when interpreting and combining heterogeneous forecasts. Understanding these issues is a prerequisite for meaningful use of the data collected in existing surveys, and it may enable design of more informative surveys.

Section 2 considers the proper interpretation of point predictions of uncertain events. Economists commonly assume that persons hold probabilistic beliefs about uncertain events. However, most surveys ask respondents for point predictions of events, not probabilistic ones. Users of point-prediction data typically do not know how respondents choose points to summarize their beliefs. This generates an unavoidable problem in analysis of the data. Other problems, which are avoidable, arise when researchers make logical errors in their interpretation of point predictions. One frequent error is to use the dispersion of point predictions across forecasters to measure the uncertainty that forecasters perceive. Another is to interpret the cross-sectional mean point prediction of a binary event as a probabilistic prediction.

Section 3 explains the simple, but under-appreciated, logical basis for a pervasive empirical finding on the performance of consensus forecasts of real-valued events. For at least a century, empirical researchers have regularly reported that the cross-sectional mean or median of a set of point predictions is more accurate than the individual predictions used to form the mean or median. However, it has only occasionally been recognized that these regularities have algebraic foundations. The one concerning mean forecasts holds whenever a convex loss function is used to measure prediction accuracy, by Jensen's inequality. The one concerning median forecasts holds whenever a unimodal loss function functions is used to measure accuracy.

Section 4 calls attention to the problem of assessing the temporal variation of forecasts made by panels of forecasters. A number of surveys periodically report the predictions of panels of professional forecasters. To study the temporal variation of forecasts, it is common to aggregate the predictions reported by panel members at each administration of the survey and analyze the time series of the aggregated predictions. Interpretation of the temporal variation in an aggregated prediction is difficult when forecasters are heterogeneous. The interpretative problem is exacerbated when panel composition changes over time.

Sections 2 through 4 examine distinct logical issues, so each section may be read in isolation from the others. Nevertheless, the entire chapter has a unifying theme in its concern with proper interpretation of heterogeneous forecasts. If forecasts were homogeneous, combining them would be trivial and the issues studied in Sections 3 and 4 would not exist. Some of the interpretative problems studied in Section 2 would persist, but others would disappear. There is ample empirical evidence that survey forecasts exhibit considerable heterogeneity. Hence, the issues treated in this section are important in practice as well as in principle.

2. Interpreting Point Predictions of Uncertain Events

Forecasters regularly give point predictions of future events. Financial analysts offer point predictions of the profit that firms will earn in the quarter ahead, macroeconomic forecasters give point predictions of GDP growth and inflation, and pundits predict who will win elections. A notable exception is that meteorologists commonly report the percent chance that it will rain during the next day. However, meteorologists give point rather than probabilistic predictions of the next day's high and low temperatures.

Thoughtful forecasters rarely think that they have perfect foresight. Hence, their point predictions can at most convey some notion of the central tendency of their beliefs, and nothing at all about the uncertainty they feel. Economists regularly assume that persons use subjective probability distributions to express uncertainty about future events. Suppose that forecasters actually have subjective probability distributions for the events they predict. Then their point predictions should somehow be related to their subjective distributions. But how?

I consider point prediction of binary outcomes in Section 2.1 and real-valued ones in Section 2.2. Section 2.3 recommends that surveys replace traditional questions seeking point predictions with ones that elicit subjective probability distributions directly.

2.1. Interpreting Point Predictions of Binary Outcomes

The idea that point predictions of binary outcomes should be related to subjective probability distributions was suggested early on by Juster (1966). Considering the case in which consumers are asked to give a point prediction of their buying intentions (buy or not buy), Juster wrote (page 664):

> "Consumers reporting that they 'intend to buy A within X months' can be thought of as saying that the probability of their purchasing A within X months is high enough so that some form of 'yes' answer is more accurate than a 'no' answer."

Thus, he hypothesized that a consumer facing a yes/no intentions question responds as would a statistician asked to make a best point prediction of a binary event. Some logical implications of

Juster's idea were later worked out in Manski (1990). I summarize part of my analysis here.

Using Point Predictions to Bound Subjective Probabilities

Suppose that a random sample of a population of persons are asked to make point predictions of a binary outcome. The outcome of interest could be a personal event such as purchase of a consumer durable, job loss, or childbirth. Or it may be a macro event such as the onset of a recession or the result of an election.

Let r and y be binary variables denoting the survey response and the subsequent outcome respectively. Thus, r = 1 if a person states that the outcome will occur and r = 0 if he states that it will not occur. Similarly, y = 1 if the outcome actually occurs and y = 0 if it does not. A researcher who observes the predictions r and the covariates x of the random sample can use the data to estimate the population distribution P(r, x). For simplicity, I will ignore statistical imprecision and assume that the researcher knows P(r, x).

Suppose that a researcher observes a person's stated prediction r and wants to learn her subjective probability that y = 1. Assume that respondents state best point predictions of the outcome, in the sense of minimizing expected loss. A best point prediction depends on the losses a respondent associates with the two possible prediction errors, (r = 0, y = 1) and (r = 1, y = 0). Whatever the loss function, the best point prediction will satisfy the condition

$$\text{(1)} \qquad \begin{aligned} r = 1 \;&\Rightarrow\; Q(y = 1|s) \;\geq\; p, \\ r = 0 \;&\Rightarrow\; Q(y = 1|s) \;\leq\; p. \end{aligned}$$

Here s denotes the information possessed by the respondent at the time that the survey question is posed, and $Q(y = 1 | s)$ denotes the subjective probability she places on outcome $y = 1$. The value of $p \in [0, 1]$ depends on the loss function that the respondent uses to form the best point prediction. This formalizes the idea originally stated by Juster (1966).

A researcher who does not know a respondent's threshold value p can conclude nothing about her subjective probability distribution. On the other hand, (1) shows that the stated prediction bounds the subjective probability given knowledge of p. In some settings, a researcher may find it credible to assume that respondents use symmetric loss functions to form their point predictions. Any symmetric loss function implies that $p = ½$.

Prediction markets are survey-like settings in which the bets placed by traders may reveal bounds on their subjective probabilities. Consider a prediction market in which traders bet on a binary outcome. Let p be the price of a bet on the event $\{y = 1\}$ and $1 - p$ be the price of a bet on $\{y = 0\}$. Thus, a dollar bet on $\{y = 1\}$ returns 1/p dollars if this event occurs and 0 otherwise, while a dollar bet on $\{y = 0\}$ returns $1/(1 - p)$ if this event occurs and 0 otherwise. If traders are risk neutral, observation that a trader bets on $\{y = 1\}$ reveals that his subjective probability for this event is at least p, while observation that the trader bets on $\{y = 0\}$ reveals that his subjective probability for $\{y = 1\}$ is no more than p. Thus, observation of the bets placed in a prediction market with risk neutral traders yield the inequalities (1). See Manski (2006) for further analysis of prediction markets, including characterization of the equilibrium price.

Using Point Predictions to Bound the Objective Probability of Personal Outcomes

Let y be a personal outcome rather than a macro event. Consider a researcher who observes the fraction $P(r = 1|x)$ of persons with covariates x who state that the outcome will occur. Suppose that the researcher wants to learn the fraction $P(y = 1|x)$ of these persons who will actually realize the outcome.

There is no necessary relationship between $P(r = 1|x)$ and $P(y = 1|x)$. The former depends on respondents' subjective probabilities for the outcome, while the latter is an objective probability. However, Manski (1990) shows that $P(r = 1|x)$ and $P(y = 1|x)$ are related if these conditions hold for all persons with covariate value x:

(a) Each person has rational expectations.

(b) Each person uses the same p to form point predictions, and the researcher knows this value.

(c) Each person's information s at the time of the survey includes her value of x.

(d) Outcome realizations are statistically independent across persons.

Given these conditions, it can be shown that knowledge of $P(r = 1|x)$ yields this sharp bound on $P(y = 1|x)$:

$$(2) \quad pP(r = 1|x) \ \leq \ P(y = 1|x) \ \leq \ pP(r = 0|x) + P(r = 1|x).$$

The width of the bound is $pP(r = 0|x) + (1 - p)P(r = 1|x)$. In the special case $p = 1/2$, the width is 1/2 regardless of the value of $P(r = 1|x)$.

Although assumptions (a) – (d) are strong, they only imply the bound on $P(y =1|x)$ given in (2). Researchers have sometimes thought that $P(r = 1|x)$ and $P(y = 1|x)$ should be much more tightly related, namely that

$$(3)\ \ P(y = 1|x) \ = \ P(r = 1|x).$$

Assumptions (a) – (d) imply only (2), not (3). Equation (3) holds if persons have perfect foresight regarding their future outcomes, in which case r always equals y. However, it generally does not hold otherwise.

Interpreting Fertility Intentions

Research in demography illustrates that misplaced faith in equation (3) can lead to misinterpretation of point predictions. Demographers have long used responses to fertility-intentions questions to predict future fertility rates. A typical such question ask a women to state (yes/no) whether she expects to give birth to a child in a specified future time period. A common practice has been to use the fraction of women who respond "yes" to predict the future fertility rate. This practice supposes that (3) holds.

Some of the literature on fertility intentions has considered deviations from (3) as evidence that women's expectations are not rational. For example, Westoff and Ryder (1977, p. 449) state:

> The question with which we began this work was whether reproductive intentions are useful for prediction. The basic finding was that 40.5 percent intended more, as of the end of 1970, and 34.0 percent had more in the subsequent five years . . . . In other words, acceptance of 1970 intentions at face value would have led to a substantial overshooting of the ultimate outcome.

That is, the authors found that $P(r = 1|x) = 0.405$, and subsequent data collection showed that $P(y = 1|x) = 0.340$. Seeking to explain the observed "overshooting," the authors state:

> one interpretation of our finding would be that the respondents failed to anticipate the extent to which the times would be unpropitious for childbearing, that they made the understandable but frequently invalid assumption that the future would resemble the present--the same kind of forecasting error that demographers have often made.

Other studies similarly presume that deviations from (3) imply women do not have rational expectations. See, for example, Davidson and Beach (1981) and O'Connell and Rogers (1983).

Equation (3) holds if women have perfect foresight about their future fertility. However, it need not hold otherwise, even if women have rational expectation. A simple example makes the point forcefully.

Suppose that women with covariates x report that they expect to have a child when childbirth is more likely than not; that is, they set $p = 1/2$. Suppose that the objective probability of having a child is 0.51 for all women and that realizations of childbirth are statistically independent across women. Then all women report that they expect to have a child, but the fraction who actually give birth is only 0.51. That is, $P(r = 1|x) = 1$ and $P(y = 1|x) = 0.51$. This large divergence between $P(r = 1|x)$ and $P(y = 1|x)$ occurs even though women have rational expectations in the example.

2.2. Interpreting Point Prediction of Real-Valued Events

Forecasters are often asked to predict real-valued outcomes such as firm profit, GDP growth, or temperature. It may be that forecasters report the means of their subjective probability distributions—their best point predictions under square loss. Or they may report their subjective medians—their best point predictions under absolute loss. However, forecasters are not specifically asked to report subjective means or medians. They are simply asked to "predict" or "forecast" the outcome.

In the absence of explicit guidance, forecasters may report different distributional features as their point predictions. Some may report subjective means, while others report subjective medians or modes. Still others, applying asymmetric loss functions, may report non-central quantiles of their subjective probability distributions. Research calling attention to and analyzing the potential heterogeneity of response practices include Elliott, Komunjer, and Timmermann (2005, 2008), Engelberg, Manski, and Williams (2009), and Clements (2009).

Interpreting Cross-Sectional Dispersion in Predictions as Forecaster Disagreement

Heterogeneous reporting practices are consequential for the interpretation of point predictions. Forecasters who hold identical probabilistic beliefs may provide different point predictions, and forecasters with dissimilar beliefs may provide identical point predictions. If so, comparison of point predictions across forecasters is problematic. Variation in predictions need not imply disagreement among forecasters, and homogeneity in predictions need not imply agreement.

Researchers in finance have sometimes interpreted cross-forecaster dispersion in point predictions to indicate disagreement in their beliefs. See, for example, Diether, Malloy, and Scherbina (2002) and Mankiw, Reis, and Wolfers (2003). If forecasters vary in the way they transform probabilistic expectations into point predictions, this interpretation confounds variation in forecaster beliefs with variation in the manner that forecasters make point predictions.

Interpreting Cross-Sectional Dispersion in Predictions as Forecaster Uncertainty

A distinct, and more severe, interpretative problem is the longstanding use of cross-sectional dispersion in point predictions to measure forecaster uncertainty about future outcomes. See, for example, Cukierman and Wachtel (1979), Levi and Makin (1979, 1980), Makin (1982), Brenner and Landskroner (1983), Hahm and Steigerwald (1999), Hayford (2000), and Giordani and Söderlind (2003). This research practice is suspect on logical grounds, even if all forecasters make their point predictions in the same way. Even in the best of circumstances, point predictions provide no information about the uncertainty that forecasters feel. This point was made forcefully over twenty years ago by Zarnowitz and Lambros (1987). Nevertheless, researchers have continued to use the dispersion in point predictions to measure forecaster uncertainty.

2.3. Probabilistic Forecasting

Even if researchers know how persons form point predictions of uncertain events, such predictions at most express the central tendency of beliefs, revealing nothing about the uncertainty persons feel. In practice, researchers usually do not know how persons form point predictions, and often misinterpret them. Surveys that presently seek point predictions would be more informative if they would instead pose probabilistic questions asking persons to reveal well-defined features of their subjective distributions.

Probabilistic questioning is especially simple when the task is prediction of a binary outcome. Instead of asking respondents to state whether the outcome will occur, they may be asked to state the percent chance that it will occur. Various methods have been used to elicit subjective distributions for continuous outcomes. Some surveys have asked respondents to state specified quantiles of their subjective distributions and others have asked respondents to state the subjective probabilities that the outcome will fall in specified intervals. If a researcher wants only to measure the central tendencies of subjective distributions, a particularly simple approach is to elicit subjective medians. One need just ask a respondent to state a value of the outcome such that there is equal probability the realization will be above or below this value.

Elicitation of probabilistic forecasts in surveys has been shown to be feasible and informative. In the realm of expert forecasting, the Survey of Professional Forecasters in the United States has long asked its panel of macroeconomists to provide probabilistic forecasts of GDP growth and inflation. These rich data were almost ignored for many years, but they are now being analyzed more regularly. See, for example, Engelberg, Manski, and Williams (2009, 2011) and Clements (2009). Similar collection of probabilistic forecasts has recently been initiated in the United Kingdom by the Bank of England. See Boero, Smith, and Wallis (2008).

In the realm of non-expert forecasting, since the early 1990s economists engaged in survey research have accumulated substantial experience with probabilistic questioning, using it to learn how broad populations perceive their futures. Manski (2004) describes the emergence of this modern field of empirical research, summarizes a spectrum of applications, and calls attention to open issues. In this space I will only give a brief description of the major American surveys, with representative citations to completed empirical studies.

Beginning in 1992 and continuing through the present, the longitudinal Health and Retirement Study (HRS) has regularly elicited probabilistic forecasts of retirement, bequests, and mortality from multiple cohorts of older Americans (Hurd and McGarry, 1995, 2002; Hurd, Smith, and Zissimopoulos, 2004). From 1994 through 2002, the nationwide Survey of Economic Expectations (SEE) asked repeated cross sections of persons to state the percent chance that they will lose their jobs, have health insurance coverage, or be victims of crime in the year ahead, and also to give their income expectations (Dominitz and Manski, 1997a, 1997b; Manski and Straub, 2000).

From 1997 on, the National Longitudinal Survey of Youth 1997 has periodically queried youth about the chance that they will become a parent, be arrested, or complete schooling in the future (Fischhoff *et al.*, 2000; Dominitz, Manski, and Fischhoff, 2001; Lochner, 2007). Probabilistic forecasts of stock market returns have been elicited in several surveys, including SEE, HRS, and the Michigan Monthly Survey (Dominitz and Manski, 2004, 2007; Hurd, 2009).

We have learned from these and other surveys that most people have little difficulty, once the concept is introduced to them, using subjective probabilities to express the likelihood they place on future events relevant to their lives. Indeed, implementation of probabilistic forecasting questions in surveys has recently spread with some success from the United States and western Europe to various developing countries with low literacy rates. See Delavande, Giné, and McKenzie (2009).

I would note in conclusion that probabilistic questioning is useful even to researchers who analyze traditional point-prediction data. We have seen that point predictions are difficult to interpret. If persons are asked to provide both a point prediction and a probabilistic forecast for an outcome of interest, then a researcher can study the relationship between the two responses. Engelberg, Manski, and Williams (2009) perform an analysis of this type, comparing the point and probabilistic responses of panel members of the Survey of Professional Forecasters.

3. The Algebra of Consensus Forecasts

For over a century, researchers studying the accuracy of forecasts have studied settings in which multiple agents are asked to give point predictions of an event and their forecasts are combined to create a *consensus forecast*. The consensus forecast is often defined to be the cross-sectional mean or median of the individual forecasts. Empirical studies have regularly found that the consensus forecast is more accurate than the individuals forecasts used to form it. Clemen (1989, p. 559) put it this way in a review article:

> "The results have been virtually unanimous: combining multiple forecasts leads to increased forecast accuracy. This has been the result whether the forecasts are judgmental or statistical, econometric or extrapolation. Furthermore, in many cases one can make dramatic performance improvements by simply averaging the forecasts."

Galton and Ox Weighing

A notable early example is Galton (1907), who presented data on the point estimates of the weight of an ox made in a weight-judging contest. He found considerable dispersion in the estimates, but the median estimate was close to the actual weight. Galton viewed this as a demonstration of the power of democracy, writing (P. 450) "According to the democratic principle of 'one vote one value,' the middlemost estimate expresses the *vox populi*, every other estimate being condemned as too low or too high by a majority of the voters." He concluded (p. 451): "This result is, I think, more creditable to the trustworthiness of a democratic judgment than might have been expected."

In a comment on Galton's work, Hooker (1907), observed that the mean estimate with the Galton data was even closer to the truth than the median was. Hooker suggests that Galton should have recommended use of the mean rather than the median as the way to combine estimates. However, Galton responded that he really meant the median. Close to a century later, Surowiecki (2004) opened his popular-science book with a summary of Galton's work, pointing to it as a leading early example of the so-called "wisdom of crowds."

Mean Forecasts and Convex Loss Functions

Researchers have been intrigued by the performance of consensus forecasts relative to individual predictions. In fact, the empirical regularity that mean forecasts perform better than the mean performance of individual forecasts follows from Jensen's inequality.

Let y_n, n = 1, . . . , N be a set of individual point forecasts of an unknown real quantity θ, let P_N denote the cross-sectional distribution of the forecasts, and let $\mu_N \equiv \int y dP_N$ denote the cross-sectional mean forecast. Let $L(\cdot, \cdot): R \times R \to [0, \infty)$ be a loss function used to measure the consequence of prediction error. Research on forecasting has typically used absolute loss $L(y, \theta) = |y - \theta|$ or square loss $L(y, \theta) = (y - \theta)^2$. When these or any other convex loss function is used, Jensen's inequality gives $L(\mu_N, \theta) \leq \int L(y, \theta)dP_N$ for all $\theta \in R$. Thus, whatever the actual value of the quantity being forecast, the loss associated with the mean forecast is no larger than the mean loss of the individual forecasts.

This simple result has long been known in statistical decision theory. There θ is a parameter to be estimated and (y_n, n = 1, . . . , N) is a *randomized estimate*, meaning that the statistician draws an integer i at random from the set (1, . . . , N) and uses y_i to estimate θ. Suppose that a convex loss function is used to measure precision of estimation. Then Jensen’s inequality implies that loss using the *non-randomized estimate* μ_N is smaller than expected loss using the randomized estimate. See Hodges and Lehmann (1950).

Research on consensus forecasts has largely disregarded the result as it has sought to explain why mean forecasts perform better than individual forecasts. A notable exception is McNees (1992), who exposited the matter clearly in the context of absolute and square loss. McNees observed that much research on forecasting did not acknowledge “these simple, well-known, yet often ignored arithmetic principles” (page 705). More recently, Larrick and Soll (2006) have elaborated on McNees’ observation, showing that experimental subjects often do not understand what they call the “averaging principle.”

Median Forecasts and Unimodal Loss Functions

McNees (1992) also recognized that the median forecast of any event must be at least as close to the truth as at least half of the individual forecasts. He wrote “it is always true that *no more* than half of the individual forecasts that define a median forecast can ever be more accurate than the median forecast.” McNees’ discussion of this algebraic truism was informal. I prove an extended result here, that holds for all quantile forecasts when L is any unimodal loss function.

Let $L(\cdot, \theta)$ be unimodal for all values of θ, with $L(\theta, \theta) = 0$ and $L(t, \theta) > 0$ for $t \neq \theta$. Let $\alpha \in (0, 1)$ and let $q_{(\alpha, N)} \equiv \inf\,[t\colon P_N(y \leq t) \geq \alpha]$ denote the α-quantile of P_N. Consider use of $q_{(\alpha, N)}$ as the consensus forecast for θ. If $q_{(\alpha, N)} = \theta$, then $P_N[L(y, \theta) \geq L(q_{(\alpha, N)}, \theta)] = P_N[L(y, \theta) \geq 0] = 1$. If $q_{(\alpha, N)} < \theta$, then $P_N[L(y, \theta) \geq L(q_{\alpha N}, \theta)] \geq P_N(y \leq q_{\alpha N}) \geq \alpha$. If $q_{\alpha N} > \theta$, then $P_N[L(y, \theta) \geq L(q_{\alpha N}, \theta)] \geq P_N(y \geq q_{\alpha N}) \geq 1 - \alpha$. Hence, $P_N[L(y, \theta) \geq L(q_{\alpha N}, \theta)] \geq \min(\alpha, 1 - \alpha)$, whatever the actual value of θ may be.

When $\alpha = ½$, this result become $P_N[L(y, \theta) \geq L(q_{(½, N)}, \theta)] \geq ½$ for all θ. In words, the loss from using $q_{(½, N)}$ to forecast θ is less than or equal to the loss from using y_n for at least half of the individual forecast values (y_n, n = 1, . . . , N). This formalizes McNees' observation, quoted above.

Discussion

The simple algebraic results derived above apply to combination of any set of forecasts, whatever their source. The forecasts could be point predictions obtained in surveys, subjective means or medians of probabilistic forecasts, or point predictions generated by econometric models.

It is important to understand that these results hold whatever the quality of the individual forecasts combined to form the consensus forecasts. The algebra shows that, whatever the truth may be, the mean forecast performs better than the mean performance of the individual forecasts and the median forecast performs better than half of the individual forecasts. However, the mean or median forecast need not perform well in an absolute sense. It may be that μ_N or $q_{(½, N)}$ is a good forecast of θ, or a terrible one.

It is also important to understand that the mean or median forecast does not outperform every individual forecast. The results show only that they outperform an individual forecast drawn at random from the available forecasts. One may have reason to think that some individual forecast is particularly credible, perhaps because this forecaster is more expert or has more information than others. In such a case, one may reasonably prefer this individual forecast to a consensus forecast.

4. Assessing the Temporal Variation of Forecasts by Panels of Forecasters

A number of surveys periodically report the macroeconomic predictions of panels of professional forecasters. Perhaps best known are the Livingston Survey and the Survey of Professional Forecasters, both presently conducted by the Federal Reserve Bank of Philadelphia. Others are the Bank of England's Survey of External Forecasters, the CESifo World Economic Survey, and the INSEE Monthly Business Survey.

To study the temporal variation of forecasts, it is common to aggregate the predictions reported by panel members at each administration of the survey and analyze the time series of the aggregated predictions. See, for example, Hafer and Hein (1985), Fair and Shiller (1989), Pennacchi (1991), Baghestani (1994, 2006), Thomas (1999), Romer and Romer (2000), Ball and Croushore (2003) and Campbell (2007). Summary reports of survey findings traditionally take this form. However, interpretation of the temporal variation in an aggregated prediction can be problematic when forecasters are heterogeneous, and is yet more difficult when panel composition changes over time.

To illustrate the interpretative problem, Engelberg, Manski, and Williams (2009b) cite this example from the quarterly Survey of Professional Forecasters (SPF). In February 2008, the Philadelphia Fed issued a release of findings from the survey administered in the first quarter of 2008, with this opening statement: "The outlook for growth in the first half of 2008 looks much weaker now than it did three months ago, according to 50 forecasters surveyed by the Federal Reserve Bank of Philadelphia. . . . . . Growth in the current quarter is projected at an annual rate of 0.7 percent, down from the projection of 2.2 percent in last year's fourth-quarter survey."

First consider the implications of heterogeneity under the assumption that the SPF panel had the same composition in the fourth quarter of 2007 (4Q2007) and the first quarter of 2008 (1Q2008). When the Philadelphia Fed reported that growth is projected at an annual rate of 0.7 percent, one cannot know whether this was a consensus across the 50 forecasters or whether they disagreed sharply in their predictions. Nor can one know whether all panel members revised their beliefs downward between 4Q2007 and 1Q2008. This type of inferential difficulty stemming from forecaster heterogeneity in a panel of fixed composition has long been recognized by researchers. See, for example, Zarnowitz and Lambros (1987), Keane and Runkle (1990), Giordani and Söderlind (2003), Pesaran and Weale (2006), and Patton and Timmerman (2010).

Engelberg, Manski, and Williams (2009b) (henceforth, EMW) call attention to the fact that further interpretative problems arise when the composition of a panel changes over time. Although the Philadelphia Fed release of findings stated that 50 forecasters participated in the survey, this actually was the number of participants in the 1Q2008 survey. The number of participants in the 4Q2007 survey was 48, of whom only 42 participated in the 1Q2008 survey. Thus, 14 forecasters participated in only one of the two surveys, 6 participating only in 4Q2007 and 8 only in 1Q2008. To an unknown extent, the weakening in beliefs about future growth reported in the release of findings could be an artifact of changing panel composition. At the extreme, 6 optimistic forecasters may have been replaced by 8 pessimistic ones.

EMW analyze the responses to SPF probabilistic questions on inflation expectations and conclude that the interpretative problem is always serious in principle and is often serious in practice. Three factors contribute to this conclusion. First, the predictions reported by SPF panel members exhibit considerable heterogeneity. Moreover, this heterogeneity exhibits strong persistence. That is, forecasters who are relatively uncertain about future inflation in one survey tend to be relatively uncertain throughout their participation in the panel. Those who expect high inflation in one survey tend to expect high inflation in other surveys. Thus, it appears that the heterogeneity observed in the SPF forecasts arises out of permanent differences between forecasters in the way that they form inflation expectations.

Second, the composition of the panel changes substantially over time, in part due to long-run turnover in the forecasters who officially serve as members of the panel and in part due to short-run variation in the panel members who actually respond to the survey. Consider, for example, the year-to-year stability of the panel. On average, 34 forecasters participated in each quarterly administration of the SPF during the 1992-2006 period. However, an average of 9 forecasters who participated in a given quarter did not participate four quarters later, with another 10 or so taking their place at that time. Thus, when comparing predictions made four quarters apart, one confronts the problem that an average of 43 or 44 forecasters participated in at least one of the two surveys, but only about 25 participated in both.

Third, little is known about the process that determines panel composition. Time-series analysis of aggregated predictions would be a well-defined inferential problem if it were credible to assume that panel members are randomly recruited from a stable population of potential forecasters and that participation in the survey after recruitment is statistically independent of forecasters' beliefs about inflation. However, evidence to justify these assumptions is not available.

The underlying difficulty is that the changing composition of the SPF panel creates a problem of partial identification due to missing data; see Manski (2007). Without knowledge of the forecaster participation process, one can only bound the distribution of inflation expectations in the panel at a given point in time, and similarly bound the distribution of changes in expectations over time. The constantly changing composition of the SPF panel implies that a large fraction of the relevant data are typically missing. Hence, the bounds are quite wide.

In the absence of knowledge of the process determining panel composition, EMW recommend against the traditional use of the time series of aggregated SPF predictions to measure the evolution of forecasters' expectations. Such time series conflate changes in the expectations of individual forecasters with changes in the composition of the SPF panel. Disentangling the two factors requires knowledge of the forecaster participation process.

Keane and Runkle (1990) also argue against using consensus forecasts. First, interpreting point forecasts as conditional expectations, they point out that the average of several expectations conditioning on different information sets need not be the conditional expectation given any one information set. Second, they argue that consensus forecasts "mask" forecaster heterogeneity. They conclude: "for both of these reasons...researchers must use individual data in order to test hypotheses about how people form expectations."

To replace analysis of aggregated predictions, EMW too recommend study of the time series of the predictions made by individual forecasters. As a prelude, we introduce a straightforward way to describe the cross-sectional heterogeneity of predictions in a given survey. We consider each forecaster separately and compute parameters that measure the central tendency and spread of the elicited subjective probability distribution for future inflation; in particular, we suggest the subjective median and interquartile range. This done, a plot showing the subjective (median, IQR) of each forecaster clearly portrays the heterogeneity of inflation forecasts at a point in time. To describe the evolution of expectations across the quarterly administrations of the survey, we recommend enhancing the plot with arrows to indicate how each forecaster changes his beliefs from one quarter to the next.

Although the EMW data analysis focuses on probabilistic inflation expectations in the SPF, the themes of the paper apply much more broadly. They apply equally well to the other probabilistic and point forecasts obtained in the SPF and, moreover, to the other panels of macroeconomic forecasters. Indeed, they apply even more broadly to panels making other types of forecasts.

References

Baghestani, H. (1994), “Evaluating Multiperiod Survey Forecasts of Real Net Exports,” *Economic Letters,* 44, 267-72.

Baghestani,H. (2006), “Federal Reserve vs. Private Forecasts of Real Net Exports,” *Economics Letters*, 91, 349-353.

Ball, L. and D. Croushore (2003), “Expectations and the Effects of Monetary Policy,” *Journal of Money, Credit and Banking* 35, 473-484.

Boero, G., J. Smith, and K. Wallis (2008), “Uncertainty and Disagreement in Economic Prediction: The Bank of England Survey of External Forecasters,” *The Economic Journal*, 118, 1107-1127.

Brenner, M. and Y. Landskroner (1983), “Inflation Uncertainties and Returns on Bonds,” *Economica*, 50, 463-468.

Campbell, S. (2007), “Macroeconomic Volatility, Predictability, and Uncertainty in the Great Moderation: Evidence from the Survey of Professional Forecasters,” *Journal of Business and Economic Statistics*, 25, 191-200.

Clemen, R. (1989), “Combining Forecasts: A Review and Annotated Bibliography,” *International Journal of Forecasting,* 5, 559-583.

Clements, M. (2009), ‘‘Internal Consistency of Survey Respondents’ Forecasts: Evidence Based on the Survey of Professional Forecasters,’’ in *The Methodology and Practice of Econometrics*, J. Castle and N. Shephard (editors), Oxford: Oxford University Press.

Cukierman, A. and P. Wachtel (1979): “Differential Inflationary Expectations and the Variability of the Rate of Inflation: Theory and Evidence,” *American Economic Review*, 69, 595-609.

Davidson, A. and L. Beach (1981), “Error Patterns in the Prediction of Fertility Behavior,” *Journal of Applied Social Psychology*, 11, 475-488.

Delavande, A., X. Giné, and D. McKenzie (2009), “Measuring Subjective Expectations in Developing Countries: A Critical Review and New Evidence,” *Journal of Development Economics*, forthcoming.

Diether, K., C. Malloy, and A. Scherbina (2002): “Differences of Opinion and the Cross Section of Stock Returns,” *The Journal of Finance*, 57, 2113-2141.

Dominitz, J. and C. Manski (1997a), “Perceptions of Economic Insecurity: Evidence from the Survey of Economic Expectations,” *Public Opinion Quarterly*, 61, 261-287.

Dominitz, J. and C. Manski (1997b), "Using Expectations Data to Study Subjective Income Expectations," *Journal of the American Statistical Association*, 92, 855-867.

Dominitz, J. and C. Manski (2004), "How Should We Measure Consumer Confidence?" *Journal of Economic Perspectives*, 18, 51-66.

Dominitz, J. and C. Manski (2007), "Expected Equity Returns and Portfolio Choice: Evidence from the Health and Retirement Study," *Journal of the European Economic Association*, 5, 369-379.

Dominitz, J, C. Manski, and B. Fischhoff (2001), "Who are Youth *At-Risk*?: Expectations Evidence in the NLSY-97," in R. Michael (editor), *Social Awakenings: Adolescents' Behavior as Adulthood Approaches*, New York: Russell Sage Foundation, 230-257.

Elliott, G., I. Komunjer, and A. Timmermann (2005). "Estimation and Testing of Forecast Rationality under Flexible Loss," *Review of Economic Studies* 72, 1107-1125.

Elliott, G., I. Komunjer, and A. Timmermann (2008), "Biases in Macroeconomic Forecasts: Irrationality or Asymmetric Loss?" *Journal of European Economic Association* 6, 122-157.

Engelberg, J. C. Manski, and J. Williams (2009), "Comparing the Point Predictions and Subjective Probability Distributions of Professional Forecasters," *Journal of Business and Economic Statistics*, 27, 30-41.

Engelberg, J. C. Manski, and J. Williams (2011), "Assessing the Temporal Variation of Macroeconomic Forecasts by a Panel of Changing Composition," *Journal of Applied Econometrics*, forthcoming.

Fair, R. and R. Shiller (1989), "The Informational Content of Ex Ante Forecasts," *The Review of Economics and Statistics*, 71, 325-331.

Fischhoff, B., A. Parker, W. de Bruin, J. Downs, C. Palmgren, R. Dawes, and C. Manski (2000), "Teen Expectations for Significant Life Events," *Public Opinion Quarterly*, 64, 189-205.

Galton, F. (1907), "Vox Populi," *Nature*, 75, 450-451.

Giordani, P. and P. Söderlind (2003), "Inflation Forecast Uncertainty," *European Economic Review*, 47, 1037-1059.

Hafer, R. and S. Hein (1985), "On the Accuracy of Time-Series, Interest Rate, and Survey Forecasts of Inflation," *Journal of Business*, 58, 377-398.

Hahm, J. and D. Steigerwald (1999), "Consumption Adjustment under Time-Varying Income Uncertainty," *Review of Economics and Statistics*, 81, 32-40.

Hayford, M. (2000), “Inflation Uncertainty, Unemployment Uncertainty, and Economic Activity,” *Journal of Macroeconomics*, 22, 315-329.

Hodges, J. and E. Lehmann (1950), “Some Problems in Minimax Point Estimation,” *Annals of Mathematical Statistics*, 21, 182-197.

Hooker, R. (1907), “Mean or Median,” *Nature*, 75, 487-488.

Hurd, M. (2009), “Subjective Probabilities in Household Surveys,” *Annual Review of Economics*, 1, 543-564.

Hurd, M. and K. McGarry (1995), “Evaluation of the Subjective Probabilities of Survival in the Health and Retirement Study,” *Journal of Human Resources*, 30, S268-S292.

Hurd, M. and K. McGarry (2002), “The Predictive Validity of Subjective Probabilities of Survival,” *The Economic Journal*, 112, 966-985.

Hurd, M., J. Smith, and J. Zissimopoulos (2004), “The Effects of Subjective Survival on Retirement and Social Security Claiming.” *Journal of Applied Econometrics,* 19, 761-775.

Juster T. (1966), “Consumer Buying Intentions and Purchase Probability: An Experiment in Survey Design,” *Journal of the American Statistical Association*, 61, 658-696.

Keane, M. and D. Runkle (1990), “Testing the Rationality of Price Forecasts: New Evidence from Panel Data,” *American Economic Review*, 80, 714-735.

Larrick, R. and J. Soll (2006), “Intuitions about Combining Opinions: Misappreciation of the Averaging Principle,” *Management Science*, 52, 111-127.

Levi, M. and J. Makin (1979), “Fisher, Phillips, Friedman and the Measured Impact of Inflation on Interest,” *Journal of Finance*, 34, 35-52.

Levi, M. and J. Makin (1980), “Inflation Uncertainty and the Phillips Curve: Some Empirical Evidence,” *American Economic Review*, 70, 1022-1027.

Lochner, L. (2007), “Individual Perceptions of the Criminal Justice System,” *American Economic Review*, 97, 440-460.

Makin, J. (1982), “Anticipated Money, Inflation Uncertainty, and Real Economic Activity,” *Review of Economics and Statistics*, 64, 126-134.

Mankiw, G. R. Reis, and J. Wolfers (2003): “Disagreement about Inflation Expectations,” NBER Macroeconomics Annual 18, Chicago: University of Chicago Press, 209-248.

Manski, C. (1990), “The Use of Intentions Data to Predict Behavior: A Best Case Analysis,” *Journal of the American Statistical Association*, 85, 934-940.

Manski, C. (2004) , “Measuring Expectations,” *Econometrica*, 72, 1329-1376.

Manski, C. (2006), “Interpreting the Predictions of Prediction Markets,” *Economic Letters*, 91, 425-429.

Manski, C. (2007), *Identification for Prediction and Decision*, Cambridge, MA: Harvard University Press.

Manski, C. and J. Straub (2000), “Worker Perceptions of Job Insecurity in the Mid-1990s: Evidence from the Survey of Economic Expectations,” *Journal of Human Resources,* 35, 447-479.

McNees, S. (1992), “The Uses and Abuses of ‘Consensus’ Forecasts,” *Journal of Forecasting*, 11, 703-710.

O’Connell, M. and C. Rogers (1983), “Assessing Cohort Birth Expectations Data from the Current Population Survey, 1971-1981,” *Demography*, 20, 369-383.

Patton, A. and A. Timmermann (2010), “Why do Forecasters Disagree? Lessons from the Term Structure of Cross-Sectional Dispersion,” *Journal of Monetary Economics*, 57, 803-820.

Pennacchi, G. (1991), “Identifying the Dynamics of Real Interest Rates and Inflation: Evidence Using Survey Data,” *Review of Financial Studies*, 4, 53-86.

Pesaran, H. and M. Weale (2006), “Survey Expectations,” in *Handbook of Economic Forecasting*, G. Elliott, C. Granger, and A. Timmermann (editors), Amsterdam: Elsevier, North-Holland.

Romer, C. and D. Romer (2000), “Federal Reserve Information and the Behavior of Interest Rates,” *American Economic Review*, 90, 429-457.

Surowiecki, J. (2004), *The Wisdom of Crowds*, New York: Random House.

Thomas, L. (1999), “Survey Measures of Expected U.S. Inflation,” *Journal of Economic Perspectives*, 13, 125-144.

Westoff, C. and N. Ryder (1977), “The Predictive Validity of Reproductive Intentions,” *Demography*, 14, 431-453.

Zarnowitz, V. and L. Lambros (1987), “Consensus and Uncertainty in Economic Prediction,” *Journal of Political Economy*, 95, 591-621.