Discussion Paper No. 1144**R**

# CALIBRATED FORECASTING AND MERGING*

by

Ehud Kalai*
Ehud Lehrer**
and
Rann Smorodinsky***

December 1995
**Revised August 1996**

*Department of Managerial Economics and Decision Sciences, J. L. Kellogg Graduate School of Management and Department of Mathematics, Northwestern University, 2001 Sheridan Road, Evanston, Illinois 60208. e-mail: kalai@casbah.acns.nwu.edu.

**Department of Managerial Economics and Decision Sciences, J. L. Kellogg Graduate School of Management and Department of Mathematics, Northwestern University, 2001 Sheridan Road, Evanston, Illinois 60208 and Raymond and Beverly Sackler Faculty of Exact Sciences, School of Mathematical Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel. e-mail: elehrer@casbah.acns.nwu.edu.

***Department of Managerial Economics and Decision Sciences, J. L. Kellogg Graduate School of Management, Northwestern University, 2001 Sheridan Road, Evanston, Illinois 60208. e-mail: rann@nwu.edu.

*The research of Kalai and Lehrer is partly supported by the National Science Foundation Economics Grant No. SBR-9223156.

Abstract

Consider a general finite-state stochastic process governed by an unknown objective probability distribution. Observing the system, a forecaster assigns subjective probabilities to future states. The resulting subjective forecast *merges* to the objective distribution if, with time, the forecasted probabilities converge to the correct (but unknown) probabilities. The forecast is *calibrated* if observed long-run empirical distributions coincide with the forecasted probabilities.

This paper links the unobserved reliability of forecasts to their observed empirical performance by demonstrating full equivalence between notions of merging and of calibration. It also indicates some implications of this equivalence for the literatures of forecasting and learning.

## 1. Introduction

The accuracy of forecasts is a central issue in economics and game theory. Of particular interest are three questions: (1) What are useful notions of being accurate, (2) What are good empirical methods to check such accuracy, and (3) What are the relationships between being accurate and performing well on empirical tests?

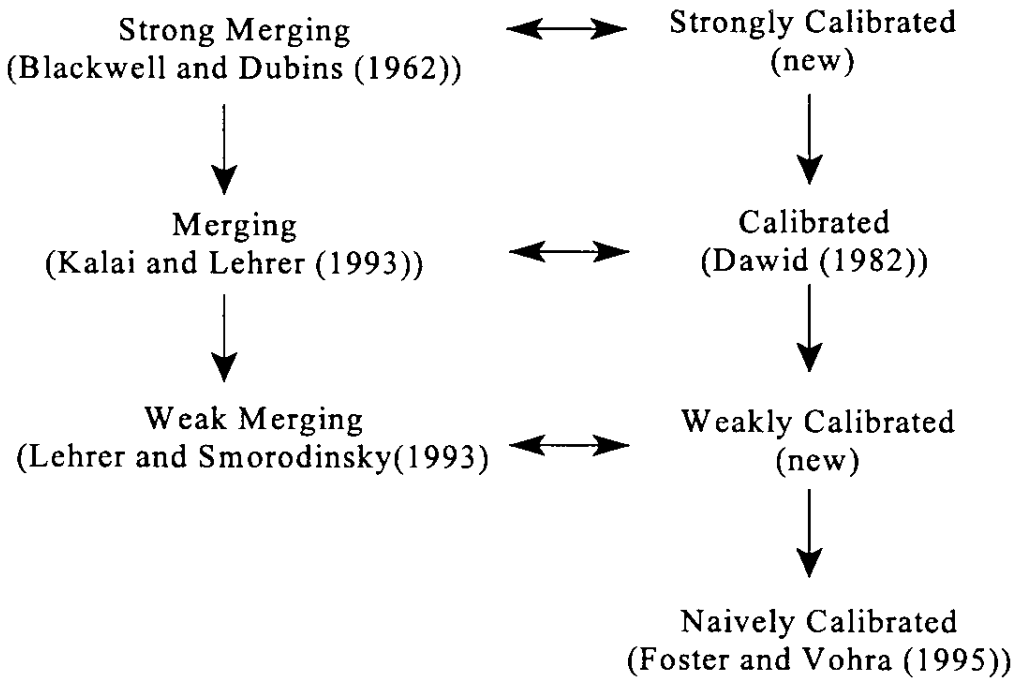
This paper studies asymptotic notions of accuracy and empirical-performance in a general finite-state stochastic process governed by an unknown, objective probability distribution. Asymptotic accuracy is achieved when a forecast merges with the truth. Originated in Blackwell and Dubins (1962), notions of merging require that for a relevant set of events, with time, the forecasted probabilities approach the real but unknown probabilities. Notions of calibration, originated in Dawid (1982), check whether the observed empirical frequencies of event-occurrences converge to their forecasted probabilities.

Recently, these notions of merging and calibration have become central to several different learning models in economics and game theory. Some of these models show that if players' forecasts merge with the true distribution or, alternatively, become calibrated with respect to the truth, then the play of the game must converge to equilibrium. Complementing these, other models describe sufficient conditions, on beliefs and behavior, that lead the players to have merging or calibrated forecasts.

This paper presents both a direct and an indirect contribution to the current literature. The direct contribution is that it establishes a full mathematical equivalence between notions of merging and calibration. That equivalence is summarized in the following diagram, which describes the logical implications between existing as well as some new notions of merging and calibration.[1]

---

[1]Our terminology is inconsistent with earlier literature where the present notions of strong merging, merging, and weak merging were named, respectively, merging, weak merging, and almost always merging. But it is consistent across notions; strong merging, merging, and weak merging correspond, respectively, to strongly calibrated, calibrated, and and weakly calibrated.




The main body of this paper consists of a formal presentation of this diagram, including definitions and proofs. The rest of this introduction and the concluding section as well will show some implications of this diagram for the forecasting and learning literature and suggest possible extensions and modifications.

The Existence of Self-Calibrating Forecasts

A forecasting method is self-calibrating if, regardless of the underlying distribution governing the system, it yields (in the long run) calibrated forecasts. Oakes (1985) illustrates the impossibility of designing such methods if full calibration is desired. Foster and Vohra (1993) show that when one employs strictly weaker notions of calibration, positive results may be obtained (see also Fudenberg and Levine (1995) and Monderer, Samet and Sela (1996) for related results).

Bayesian forecasts--in which a subjective prior over a set of possible objective distributions is updated as observations become available--have been shown to merge regardless of the underlying objective distribution, whenever this distribution is in some sense compatible with prior beliefs (see, e.g., Blackwell and Dubins (1962), Kalai and Lehrer (1993), Lehrer and Smorodinsky (1993, 1995)). As a result of the equivalence relations of merging with calibration demonstrated in this paper, it follows that calibration (at various levels) is also achieved. Thus, at least within the restricted Bayesian environments, Bayesian forecasts are self calibrating.

Conjectural, Self Confirming Subjective Equilibria

Conjectural, self confirming, and subjective equilibria--*CSS* for short--have been identified repeatedly in recent learning models as more natural than their established, objective counterparts. For a discussion of conjectural equilibria, see Hahn (1973) and Battigalli et al. (1992 )); for self confirming, see Fudenberg and Kreps (1988) and Fudenberg and Levine (1993)); and for subjective equilibria, see Kalai and Lehrer (1993) and (1994). The idea, going back to Hayek (1937), is that such equilibria should satisfy two properties.

The first property, subjective optimization, posits that a player operating in an uncertain, dynamic environment, will make subjective assumptions and assessments about how his environment works, e.g., about its unknown parameters and transition rules. He will then choose his own strategy to be optimal relative to his own subjective assessments.

However, if a player's assumptions are contradicted by events he observes or by the frequency with which they occur, he will revise his beliefs and his strategy. Thus, for equilibrium to prevail, a second condition is required--that of confirming individual beliefs.

The current literature on CSS equilibria formulates subjective optimization by assuming that players maximize expected utility in the Bayesian manner--that they hold subjective prior probabilities over unknown data and act optimally relative to these priors. Formulations of the property of belief confirmation require that the subjective probabilities assigned by each player

to his own observable events coincide with the real objective probabilities determined by nature and by all players' actions. In other words, each player's subjective forecast of own observable events should be accurate: in the language of this paper, his forecast should merge to the true distribution. But since he does not know the true distribution all he can do to test his beliefs is check whether the events he observes, and the frequency with which they occur, are consistent with his subjective assessments; or in the language of this paper, he can only check whether his beliefs are calibrated.

This observation suggests that a better definition of belief confirmation would require players' beliefs to be calibrated rather than accurate. With this in mind, the equivalence relations of merging and being calibrated presented in this paper are fundamental if we are to justify current definitions of CSS equilibria. Also, the different levels of calibration introduced herein could help establish more refined levels of belief confirmation, and thus of CSS equilibria. While such equivalence relations and other subtleties may be trivial in i.i.d. generated environments--e.g., in the repeated play of an equilibrium of a finite horizon game--they are tremendously important if we are to understand these concepts in nonstationary environments, e.g., when players learn the behavior of others in an infinite game.

The repeated-game learning model of Foster and Vohra, discussed below, is one example. Their players act optimally against beliefs which are calibrated in the weakest naive sense. Thus, they may be viewed as playing a weak version of CSS equilibrium. The result of their convergence to correlated equilibrium, described below, can therefore be reinterpreted as a general result about such CSS equilibria. Namely, the cumulative empirical play of naively-calibrated CSS equilibria approximates a correlated equilibrium of the stage game.

Connections Among Different Learning Models

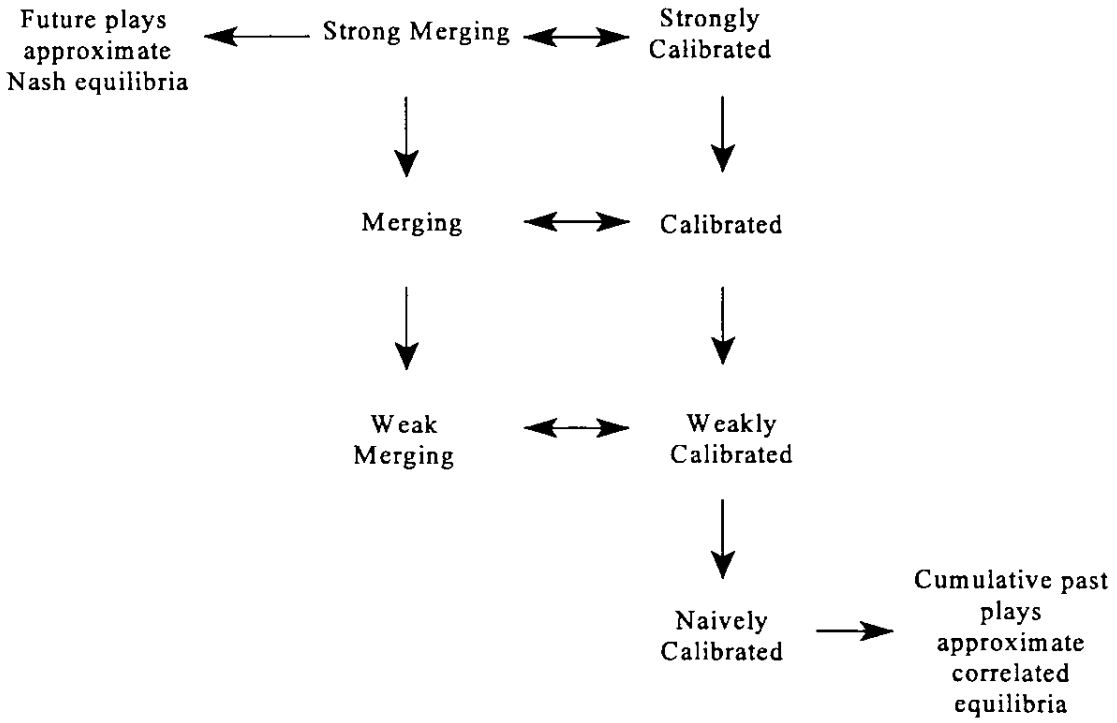
As we have already mentioned, recent learning models have shown convergence to equilibrium by players who optimize with respect to subjective forecasts that merge or become

calibrated. The logical implications between the notions of merging and calibration presented in this paper establish relationships among such convergence results.

For a brief illustration, we restrict ourselves to the case of two-player learning in an infinitely repeated game. We assume that the two players know the stage game--at least their own portions of the payoff table--but do not know the repeated game strategies their opponents choose. The players hold probabilistic assessments regarding the stage actions of their opponents, and their own strategies are optimal against these assessments. Each player's assessment, together with the knowledge of his own strategy, will result in a forecast about the play path of the repeated game. We consider two types of results.

Under strong merging, Kalai and Lehrer (1993) show that, for any discount parameter, after sufficiently long time T, the distributions over future play paths approximate Nash equilibria play of the repeated game. Under naive calibration, Foster and Vohra (1995) show that, for myopic players, after sufficiently long time T, the cumulative empirical distributions of past plays will approximate correlated equilibria of the one shot game.

Incorporating these results into the previous diagram, we obtain:




Notice that by using the above diagram we may now obtain direct learning results. For example, if the players' forecasts strongly merge, merge, or even weakly merge, they must be naively calibrated. And through the Foster and Vohra result (the diagram's bottom-right side implication), we may deduce that the frequencies of empirical play must eventually approximate correlated equilibria of the stage game. Thus, the cumulative play of optimizers who use merging forecasts will approximate correlated equilibria of their stage game.

But, ignoring the use of the diagram, stronger conclusions hold. Consider, for example, myopic players with strongly merging forecasts. By the results of Kalai and Lehrer we know that their stage game plays will eventually approximate Nash equilibria of the stage game (myopic players should be viewed in the Kalai and Lehrer model as ones with very small discount parameter). Thus, their cumulative play will eventually be near the set of convex combinations of Nash equilibria, a set strictly smaller than the one consisting of correlated

equilibria.

The fact that the result obtained through the diagram is weaker than possible suggests that there is room for refining the diagram further. It would also be useful to study extensions that cover other behavioral assumptions. For example, can the above diagram be tied to the new result of Hart and Mas-Colell (1996), about empirical convergence to correlated equilibrium by regret-minimizing players?

## 2. Forecasting Rules

Consider a finite set of states, $\Omega$, and a stochastic process selecting one state $\omega_t \in \Omega$ in each time period, $t = 1,2,\ldots$. For each outcome $\omega^\infty = (\omega_1,\omega_2,\ldots) \in \Omega^\infty$ we let $\omega^t = (\omega_1,\ldots,\omega_t)$ denote the history of length t. $\Omega^\infty$ is naturally endowed with the $\sigma$-algebra $\mathcal{F}$ generated by all histories. We let $\mu$ denote the probability distribution on $\Omega^\infty$ governing this process, and thus $\mu(\omega_{t+1}|\omega^t)$ denotes the conditional probability that the next state will be $\omega_{t+1}$ given the history $\omega^t$.

A forecasting rule, $\tilde{\mu}$, assigns subjective assessments to such conditional probabilities. Thus, $\tilde{\mu}(\omega_{t+1}|\omega^t)$ denotes the probability that the forecaster assigns to the next state being $\omega_{t+1}$ after observing the history $\omega^t$. Assuming that for every history $\omega^t$ $\tilde{\mu}(\bullet|\omega^t)$ is a probability distribution over $\Omega$, $\tilde{\mu}$ induces another unique well-defined distribution on $\Omega^\infty$. We refer to $\mu$ and $\tilde{\mu}$, respectively, as the real and the subjective distributions. (Our temporary abuse of notation is justified, since the conditional probabilities determined by the induced $\tilde{\mu}$ are indeed the subjective probabilities from which $\tilde{\mu}$ is constructed. See the concluding section for additional comments.)

In weather prediction, for example, $\Omega = \{0,1\}$ may denote, respectively, states of a dry or rainy day. $\tilde{\mu}(1|\omega^t)$ denotes the assessed probability of rain on the next day after observing the t-period history of rain described by $\omega^t$. The real probability of rain, however, is $\mu(1|\omega^t)$. In a repeated-game application, $\omega_t$ describes the t-period vector of actions taken

by n-players, $\tilde{\mu}(\omega_{t+1}|\omega^t)$ describes a player's forecasted probability of the action vector $\omega_{t+1}$ after the history of plays $\omega^t$.

## 3. Naive-Calibration

To test empirically the reliability of a forecast, Dawid (1982) introduced a notion of being calibrated where the observed frequencies of events match the probabilities forecast for them. For a simple illustration of this idea, we will first discuss a substantially weaker notion of calibration, defined in the spirit of Foster and Vohra (1995). (For Foster and Vohra, being calibrated is an assumption that leads to equilibrium. Thus, the use of a weaker notion is desirable, since it results in a stronger theorem.)

Definition 1: $\tilde{\mu}$ is naively-calibrated with $\mu$ if $\mu$-almost every $\omega^\infty \in \Omega^\infty$ satisfies the following condition. For every state $s \in \Omega$ and a number $0 \le p \le 1$, if $\Sigma_{t=1}^{\infty} I(\tilde{\mu}(s|\omega_{t-1}) = p) = \infty$, then

$$(1) \qquad \lim_{T\to\infty} \frac{\sum_{t=1}^{T} I(\tilde{\mu}(s|\omega^{t-1}) = p)I(\omega_t = s)}{\sum_{t=1}^{T} I(\tilde{\mu}(s|\omega^{t-1}) = p)} = p.$$

So, naive-calibration requires that for all the times that a forecaster says that the next state will be s with probability p, the long-run empirical frequency of such occurrences will indeed be p. However, Foster and Vohra (1995) provide an example of $\tilde{\mu}$ which is naively-calibrated with $\mu$ and yet lacks any predicting power.

Example 1: Let $\Omega = \{0,1\}$. Suppose that $\mu$ assigns only one sequence in $\Omega^\infty$ a positive probability. Specifically, let $\mu((0,1,0,1,\ldots)) = 1$. A constant forecast of (1/2,1/2), i.e., a probability of ½ is assigned to both 0 and 1 after every history, is certainly naively-calibrated

with $\mu$.

Example 1 illustrates a gap between naive-calibration and accurate prediction. While predicting the next state in the 0-1 alternating pattern seems easy, the constant forecast completely fails in this task, yet it is declared naively-calibrated. Notice, however, that this poor power to predict can be detected empirically--for example, if we compare the forecasted probabilities of 1's made prior to even periods, which is always ½, with the empirical frequencies of 1's on even periods, which is 100 percent.

### 3.1 Checking Rules and Calibration

Generalizing the idea of checking only at even times, this section develops a general notion of a checking rule, and explains what it means for a forecast to pass it. We then distinguish levels of calibration of forecasts according to the size of the sets of checking rules they pass.

Example 2: Again, let $\Omega = \{0,1\}$ but this time let $\mu$ be defined through the following Markovian dynamics: the next state is the same as the current state with probability 0.99, and the other state with probability 0.01. Specifically, let the initial state be 0, $\mu(\omega_1 = 0) = 1$, and for any history $\omega^t$, $\mu(s|\omega^t) = .99$ if $s = \omega_t$ and $\mu(s|\omega^t) = 0.01$ if $s \neq \omega_t$.

In Example 2, states change infrequently, but the times of change are stochastic. A good forecaster should be able to predict that a state is not likely to change from one period to the next. Yet, the constant forecast, (.5,.5), is still naively-calibrated. The inaccuracy of this forecast will be detected, however, if we put it to a test only at the (random) times when the last observed state is 1. Thus, the decision about when to check should allow dependence on the history of observed states. Extending this logic to cover forward-looking patterns leads to

the following definition.

Definition 2: A checking rule consists of two functions, C and D, both defined on the domain of all histories, $\cup_{t=0}^{\infty} \Omega^t$ ($\Omega^0$ is a singleton set containing the empty history). For every history $\omega^t$, $C(\omega^t) \in \{0,1\}$, and $D(\omega^t)$ is an event measurable at a finite time $t + s$ where s is a nonnegative integer that may depend on $\omega^t$. (An event is measurable at time r if it is in the $\sigma$-field determined by cylinders of length r, i.e., all the outcomes with a common initial history $\omega^r$ belong to the event or they all do not belong to the event).

The interpretation of (C,D) is the following. The function C indicates whether after the history $\omega^t$ an inspection must take place. In case an inspection takes place, D determines what event is to be inspected. For example, let $\omega^2$ = (rain,rain), $C(\omega^2) = 1$, and $D(\omega^2)$ = {(rain,rain,sun),(rain,rain,rain,sun)} (formally, $D(\omega^2)$ contains all the outcomes $\omega^\infty$ with $\omega^3$ = (rain,rain,sun) or $\omega^4$ = (rain,rain,rain,sun), and it is measurable at time 4). This means that if the first two days are rainy, we will check the event that the weather will change some time over the next two days.

To determine whether a forecast passes the checking rule (C,D) at the outcome $\omega^\infty = (\omega_1,\omega_2,\ldots)$ we will study the long-run rate of occurrence of the events $D(\omega^{t_1}),D(\omega^{t_2}),\ldots$ along the subsequence of histories $(\omega^{t_i})$ whose C value is one. This rate should match the long-run average of the forecasted probabilities for these events.

Two important comments on the definition of C and D need to be made here. First, requiring that $D(\omega^t)$ be a measurable in finite time is crucial to our ability to determine eventually whether it occurred or not. However, the length of all $D(\omega^t)$ is not bounded as we vary $\omega^t$. Thus, forecasters with accurate unbounded long-run forecasts can be identified.

Second, C and D being only functions of $\omega^t$ seems to exclude checking rules which depend on earlier forecasted values. However, this exclusion is only artificial since we implicitly assume that the checker knows the forecasting rule. Thus, all values forecast at

times 0 through t, being functions of $\omega^t$, can play a part in the checking decision at time t. If the checker does not know the forecasting rule, we should expand the arguments of C and D to depend also on the history of the forecasted values.

We may now describe naive calibration by a restricted set of checking rules. Define for every $0 \leq p \leq 1$ and $s \in \Omega$ the $(\tilde{\mu},p,s)$-checking rule as $C(\omega^t) = 1$ if $\tilde{\mu}(s|\omega^t) = p$ and 0 otherwise. And let $D(\omega^t)$ consist of the event $(\omega^t,s)$, the concatenation of $\omega^t$ with s. Notice that (1) may be written as

$$\lim_{T\to\infty} \frac{\sum_{t=1}^{T} C(\omega^{t-1})I(\omega_t = s)}{\sum_{t=1}^{T} C(\omega^{t-1})} = \lim_{T\to\infty} \frac{\sum_{t=1}^{T} \tilde{\mu}(s|\omega^{t-1})C(\omega^{t-1})}{\sum_{t=1}^{T} C(\omega^{t-1})},$$

with the right side actually being the constant p. This inspires the following general definition.

Definition 3: We say that $\tilde{\mu}$ passes the checking rule (C,D) if for $\mu$ almost every $\omega^\infty = (\omega_1,\omega_2,\ldots)$, whenever $\Sigma_{t=0}^{\infty} C(\omega^t) = \infty$

$$\lim_{T\to\infty} \frac{\sum_{t=0}^{T} C(\omega^t)\left[I(\omega^\infty \in D(\omega^t)) - \tilde{\mu}(D(\omega^t)|\omega^t)\right]}{\sum_{t=0}^{T} C(\omega^t)} = 0. \tag{2}$$

Notice that, by definition, $\tilde{\mu}$ is naively-calibrated with $\mu$ if $\tilde{\mu}$ passes all $(\tilde{\mu},p,s)$-checking rules where $0 \leq p \leq 1$ and $s \in \Omega$. However, in addition to the improvements already discussed, the new definition enables us to check the forecaster in other important cases not covered by naive-calibration. Suppose, for example, that along the outcome $\omega^\infty$ a state s is

forecasted with distinct subjective probabilities, say 0.91,0.901,0.9001,.... Then (along $\omega^*$) $\tilde{\mu}$ vacuously passes the $(\tilde{\mu},p,s)$-checking rules for every value of p, since the value p is forecast at most once. Thus, testing for naive-calibration is meaningless in such a case. On the other hand, under Definition 3, a checking rule that depends only on $\tilde{\mu}$ and s is allowable. It will enable us to check (along $\omega^*$) that the long-run rate of occurrence of s equals its long-run average forecasted probability of 0.90.

Motivated by the various improvements of the naively calibrated subjective measure in Example 1, and using the notion of checking rules, we define a partial order over subjective measures.

Definition 4: Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ be two subjective distributions. $\tilde{\mu}_1$ is better calibrated than $\tilde{\mu}_2$ (with respect to $\mu$) if $\tilde{\mu}_1$ passes every checking rule that $\tilde{\mu}_2$ passes.

Example 1 (Revisited): Let $\mu$ and $\tilde{\mu}$ be as in Example 1. Let $\tilde{\mu}_2$ be a subjective distribution defined by $\tilde{\mu}_2(1|\omega^{t-1}) = ½$ if 7 divides t but $\tilde{\mu}_2(\omega_t|\omega^{t-1}) = 1$ for t not divisible by 7 (the forecaster tells the truth on weekdays but randomizes every Sabbath). We can easily see that $\tilde{\mu}_2$ is better calibrated than the constant (.5,.5) forecasting rule $\tilde{\mu}$.

We proceed to identify natural levels of calibration by restricting ourselves to natural classes of checking rules. We start with maximal calibration.

Definition 5: $\tilde{\mu}$ is strongly calibrated with $\mu$ if it passes all checking rules.

Dawid (1985) restricted himself to checking rules that consider only next-period events. The resulting notion of calibration can be seen to be identical to those obtained by checking rules restricted to uniformly bounded finite future-horizons (for some $s \geq 1$ for all $\omega^t$, $D(\omega^t)$ is

an event measurable at time t + s). We provide the simpler definition.

Definition 6: (C,D) is a short-run checking rule if for every $\omega^t$, $D(\omega^t)$ is an event measurable at time t + 1.

Definition 7 (Dawid (1985)): $\tilde{\mu}$ is calibrated with $\mu$ if it passes all short-run checking rules.

It is natural also to pose restrictions on checking rules through the function C, i.e., the frequency of checking.

Definition 8: A checking rule (C,D) is attentive if, for almost every $\omega^\infty$, $\liminf_{T\to\infty} (1/T) \sum_{t=1}^{T} C(\omega^t) > 0$. And, in particular, a checking rule is full if $C(\omega^t) = 1$ for all $\omega^t$.

Definition 9: $\tilde{\mu}$ is weakly calibrated with $\mu$ if it passes all short-run attentive checking rules.

## 4. Merging and Calibration

Notions of calibration check the consistency of the forecasted probabilities with empirically observed frequencies. We now switch to notions of merging, where subjective probabilities that are forecast are required to converge to the true unknown objective probabilities.

Definition 10 (Blackwell and Dubins (1962)): The subjective distribution $\tilde{\mu}$ strongly merges to $\mu$ if for $\mu$-almost every $\omega^\infty = (\omega_1,\omega_2,\ldots)$

$$\sup_{A\in\mathcal{F}} \left|\tilde{\mu}(A\,|\,\omega^t) - \mu(A\,|\,\omega^t)\right| \underset{t\to\infty}{\to} 0.$$

Definition 11 (Kalai and Lehrer (1994)): The subjective distribution $\tilde{\mu}$ merges to $\mu$ if for $\mu$ almost every $\omega^\infty = (\omega_1,\omega_2,\ldots)$

$$\sup_{A\subset\Omega} |\tilde{\mu}(A\,|\,\omega^t) - \mu(A\,|\,\omega^t)| \to_{t\to\infty} 0.$$

(The abused notation $\mu(A\,|\,\omega^t)$ stands for the probability of the event A occurring at time t + 1 given the history $\omega^t$.) When $\tilde{\mu}$ merges with $\mu$ the conditional probabilities of next period events--computed by $\tilde{\mu}$ and $\mu$--become arbitrarily close to each other. Under strong merging, this holds true not just for short-run events, but also for all measurable events $A \in \mathcal{F}$, even infinite horizon ones.

Definition 12 (Lehrer and Smorodinsky (1993a)): $\tilde{\mu}$ weakly merges to $\mu$ if for $\mu$ almost every $\omega^\infty = (\omega_1,\omega_2,\ldots)$

$$\sup_{A\subset\Omega} |\tilde{\mu}(A\,|\,\omega^t) - \mu(A\,|\,\omega^t)| \to_{t\to\infty} 0.$$

where t converges to infinity along some (random) subsequence of times L of density 1, i.e., $\liminf_n (\{1,\ldots n\} \cap L)/n = 1$.

Definition 12 differs from Definition 11 in that it allows for a discrepancy between $\tilde{\mu}(A|\omega^t)$ and $\mu(A|\omega^t)$ along a sparse set of periods t--namely, along a set whose complement has density 1.

We now turn to the connections between merging and calibration. Dawid (1985) proved that $\mu$ is calibrated with itself, but his proof actually shows that if $\tilde{\mu}$ merges to $\mu$, then $\tilde{\mu}$ is calibrated with $\mu$. In fact, both are equivalent, as stated in the following theorem. For

comprehensiveness, we include Dawid's proof.

Theorem 1: $\tilde{\mu}$ merges to $\mu$ if and only if $\tilde{\mu}$ is calibrated with $\mu$.

Proof: We first prove that calibration implies merging. Define for every number $d > 0$, $B \subset \Omega$ and stage t:

$$C^{d,B}(\omega^t) = 1, \qquad \text{if } \tilde{\mu}(B|\omega^t) - \mu(B|\omega^t) > d,$$

$$C^{d,B}(\omega^t) = 0, \qquad \text{otherwise.}$$

If $\tilde{\mu}$ does not merge with $\mu$, then (without loss of generality) there is an event $A \subset \Omega^\infty$, $\mu(A) > 0$, s.t. for every $\omega^\infty \in A$ there is an infinite sequence $B^t = B^t(\omega^\infty) \subset \Omega$ for which $C(\omega^t) = C^{d,B^t}(\omega^t) = 1$. Define for all such $\omega^t$ $D(\omega^t) = B^t(\omega^\infty)$.

We prove now that $\tilde{\mu}$ fails the (C,D)-calibration test.

Define the random variables

$$X_T(\omega^\infty) = I(\omega_T \in B^T) \quad \text{and} \quad Y_T = \frac{X_T - E(X_T|\omega^{T-1})}{\sum_{t=1}^{T} C(\omega^{t-1})} C(\omega^{T-1}),$$

where 0/0 is defined to be 0. Recall that $C(\omega^{T-1})$ is either 0 or 1 and therefore $Y_T$ is either $\frac{X_T - E(X_T|\omega^{T-1})}{\sum_{t=1}^{T} C(\omega^{t-1})}$ or 0. In both cases $E(Y_T|\omega^{T-1}) = 0$. Therefore, $Y_T$ is a Martingale difference (see Shiryayev (1984), p. 453); hence, $S_T = \sum_{t=1}^{T} Y_T$ is a Martingale. In order to apply the Martingale convergence theorem it is sufficient to show that the second moment of $S_T$ is uniformly bounded:

$$E(Y_T^2) \le \frac{\left[C(\omega^{T-1})\right]^2}{\left[\sum_{t=1}^{T} C(\omega^{t-1})\right]^2} \cdot E\left[(X_T - E(X_T|\omega^{T-1}))^2\right] \le \frac{1}{4} \cdot \frac{\left[C(\omega^{T-1})\right]^2}{\left[\sum_{t=1}^{T} C(\omega^{t-1})\right]^2}.$$

Therefore:

$$E(S_K^2) = \sum_{T=1}^{K} E(Y_T^2) \le \frac{1}{4} \cdot \sum_{T=1}^{K} \frac{C(\omega^{T-1})^2}{\left[\sum_{t=1}^{T} C(\omega^{t-1})\right]^2} \le \frac{1}{4} \cdot \sum_{T=1}^{\infty} \frac{C(\omega^{T-1})^2}{\left[\sum_{t=1}^{T} C(\omega^{t-1})\right]^2} = \frac{1}{4} \cdot \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{24}.$$

By the Martingale convergence theorem (again, see Shiryayev (1984), p. 476), $S_T$ converges almost surely. By Kronecker's lemma (see Feller (1971)), if $\sum_{t=0}^{\infty} C(\omega^t) = \infty$ and if $\sum_{\tau=1}^{T} \frac{X_t - E(X_t|\omega^{t-1})}{\sum_{\tau=1} C(\omega^{\tau-1})} C(\omega^{t-1}) \underset{t\to\infty}{\to} 0.$ converges, then

$$(3) \qquad \frac{\sum_{t=0}^{T} \left[X_t - E(X_t|\omega^t)\right] C(\omega^t)}{\sum_{t=0}^{T} C(\omega^t)} \underset{T\to\infty}{\to} 0, \ \mu\text{-a.e.}$$

By the definition of C, whenever $C(\omega^{t-1}) = 1$, $\tilde{\mu}(B^t|\omega^{t-1}) - \mu(B^t|\omega^{t-1})$ which is equal to $\tilde{\mu}(B^t|\omega^{t-1}) - E(X_t|\omega^{t-1})$ is greater than d. Thus, (3) implies

$$(4) \qquad \frac{\sum_{t=0}^{T} \left[X_{t-1} - \tilde{\mu}(B^{t-1}|\omega^t)\right] C(\omega^t)}{\sum_{t=0}^{T} C(\omega^t)} \le \frac{\sum_{t=0}^{T} \left[X_{t-1} - E(X_{t-1}|\omega^t)\right] C(\omega^t)}{\sum_{t=0}^{T} C(\omega^{t-1})} - \frac{\sum_{t=0}^{T} d \cdot C(\omega^t)}{\sum_{t=0}^{T} C(\omega^t)} \underset{t\to\infty}{\to} 0 - d.$$

In other words, the left side of (4) is asymptotically bounded by $-d < 0$, meaning that $\tilde{\mu}$ fails the C-calibration test on an event A having $\mu$-positive probability. This contradicts the

assumption about $\tilde{\mu}$. We conclude, therefore, that $\tilde{\mu}$ merges with $\mu$.

As for the converse, assume that $\tilde{\mu}$ merges with $\mu$. Fix a checking rule, (C,D), where $D(\omega^t) = (\omega^t, B^t)$, where $B^t \subset \Omega$. (Thus, (C,D) is a short-run checking rule.) We obtain from the previous arguments that

$$(5) \qquad \frac{\sum_{t=0}^{T} \left[X_{t+1} - \mu(D(\omega^t)|\omega^t)\right] C(\omega^t)}{\sum_{t=0}^{T} C(\omega^t)} \underset{T\to\infty}{\to} 0, \ \mu\text{-a.e.}$$

Moreover, by merging, the difference between $\mu(D(\omega^t)|\omega^t)$ and $\tilde{\mu}(D(\omega^t)|\omega^t)$ converges to 0. Therefore, replacing $\mu(D(\omega^t)|\omega^t)$ by $\tilde{\mu}(D(\omega^t)|\omega^t)$ in (5), we obtain

$$\lim_{T\to\infty} \frac{\sum_{t=0}^{T} \left[X_t - \tilde{\mu}(D(\omega^t)|\omega^t)\right] C(\omega^t)}{\sum_{t=0}^{T} C(\omega^t)} = 0.$$

Therefore, $\tilde{\mu}$ passes the (C,D)-calibration test. As this is arbitrary, $\tilde{\mu}$ is calibrated. ■

So far we have obtained an equivalence between merging and calibration. We proceed now to the equivalence between the strong counterparts.

Theorem 2: $\tilde{\mu}$ strongly merges to $\mu$ if and only if $\tilde{\mu}$ is strongly calibrated with $\mu$.

The proof for Theorem 2 involves two main ideas. The first is that of Theorem 1--namely, the use of the Martingale convergence theorem. The second idea is based on the fact that the algebra generated by the set of all cylinders is a basis for the $\sigma$-algebra on $\Omega^\infty$. This

idea is captured by the following lemma:

Lemma 1 (Araujo and Sandroni (1994)): If there exists a set A with $\mu(A) > 0$ and $\tilde{\mu}(A) = 0$, then there exist events $\{B^t\}_{t=1}^{\infty}$ in the algebra generated by all histories such that $\tilde{\mu}(B^t) \to 0$, while $\mu(B^t) > (\frac{1}{2})\mu(A) > 0$ for all t.

Proof of Theorem 2: We begin by showing that strong calibration implies strong merging. We assume $\tilde{\mu}$ is strongly calibrated with $\mu$ and we show that $\tilde{\mu}$ strongly merges to $\mu$. Suppose this is not the case. Then by the Blackwell-Dubins' (1962) theorem, there is a set, $A \subset \Omega^{\infty}$, such that $\mu(A) > 0$ and $\tilde{\mu}(A) = 0$. We conclude that $\tilde{\mu}(A|\omega^t) = 0$ for all $\omega$ and all t. By Levy's 0-1 law, for almost all $\omega^{\infty} \in A$, $\mu(A|\omega^t) \to 1$, and therefore $\mu(A|\omega^t)$ can be uniformly bounded from below, from some (random) stage on by some $c > 0$.

By Lemma 1, there exists a sequence of sets in the algebra generated by the set of all cylinders, $\{B^{t,r}(\omega^t)\}_{r=1}^{\infty}$ s.t. $|\tilde{\mu}(B^{t,r}(\omega^t)|\omega^t)| < 1/r$, and $\mu(B^{t,r}(\omega^t|\omega^t) > (\frac{1}{2})c > 0$. We focus on the sequence $\{B^{t,t}\}_{t=1}^{\infty}$. Consider the checking rule defined by $C(\omega^t) = 1$ for all $\omega$ and t and $D(\omega^t) = B^{t,t}(\omega^t)$ if $\omega^t$ can be extended to some $\omega^{\infty} \in A$, and let $C(\omega^t)$, $D(\omega^t)$ be arbitrary otherwise. By similar arguments to those of Theorem 1, one can show that $\tilde{\mu}$ does not pass the (C,D)-calibration rule on the set A and, therefore, the desired contradiction is reached.

As for the second part, we need to show that strong merging implies strong calibration. Because the proof repeats the same two ideas, we leave it to the reader. ■

A similar counterpart is obtained between weak calibration and weak merging.

Theorem 3: $\tilde{\mu}$ weakly merges to $\mu$ iff $\tilde{\mu}$ is weakly calibrated with $\mu$.

Proof: We omit the proof since it is similar to the proof of Theorem 1. ■

## 5. Concluding Remarks

### A Local Definition of Calibration

The order of quantifiers used in the definition of calibration is one of two seemingly natural choices. Recall that being calibrated requires passing all the checking rules in a given set of checking rules, and to pass a given checking rule means passing it at almost all outcomes . Thus, being calibrated is not a local property--i.e., defined first at an outcome and then for all outcomes.

The seemingly more attractive alternative is to have a local definition. First define being calibrated at an outcome as passing all the checking rules at this outcome, and then define being calibrated as being calibrated at almost all outcomes. The attractiveness of a local definition is that it may be checked locally, i.e., one can actually determine whether one is calibrated at the realized outcome.

Similarly with Dawid (1982), we follow the less appealing global definition. This is in part because with a local definition it is impossible to obtain full calibration. Consider even the extreme case of a perfect forecast, where the forecasted probabilities are exactly correct. While for any given checking rule there is a measure one set of outcomes that pass, therefore implying global calibration , it is the case that for every given outcome there are many checking rules that fail. Thus at no outcome will the perfect forecast be locally calibrated.

This observation leads to questions regarding a proper definition of local calibration. It may be possible to obtain a meaningful local definition of calibration by other natural restrictions of the set of checking rules. For example, if one considers only a countable set of checking rules then a global definition is equivalent to a local one. Alternatively, if we restrict ourselves to naive checking rules, we can have a measure one set of outcomes with the property that the forecast passes all the naive checking rules at each and every outcome in that set. In other words, naive calibration may be defined locally. There are other possible ways, and other reasons, to restrict the set of checking rules, as we discuss next.

Restricting the Sets of Checking Rules

All three notions of calibration discussed in this paper require passing an infinite number of checking rules. This again strengthens the results that merging implies calibration. But, since it is difficult to impose infinitely many tests on a forecast, the other direction of the equivalences--that calibration implies merging--becomes less useful. Hence, one wishes to study several questions. For example, for a class of checking rules C, find a sufficient subset of rules R such that a forecast passing all the checking rules in R will also pass all the checking rules in C.

To check weak calibration, we can show that the set of full short-run checking rules (checking in every period) suffices for the class of attentive short-run checking rules. While this presents a substantial reduction (in the number of checking rules), the set is still infinite.

In some situations, however, a finite number of checking rules will suffice. Consider, for example, the case that $\mu$ is Markovian with a finite set of states. For every ordered pair of states (s,r), one can construct a small and finite number of checking rules to check the average forecasted probabilities of transitions from the state s to r and to be compared to the actual empirical rate of such transitions. Moreover, one does not have to know the real transition probability in constructing these checking rules; all one has to know is that for every pair of states the transition probability is fixed.

This observation suggests that we enlarge the notion of sufficiency of R for C to be relative to a known set of possible true distributions, P, from which the unknown distribution $\mu$ was drawn. For example, following the previous discussion, if we select P to be the set of Markovian probability distributions on a fixed and finite set of states, we should also be able to find a finite set of checking rules sufficient to check calibration, and, hence, also sufficient for merging.

While the Markovian case does not apply to nonstationary behaviors--e.g., cases from game theory--it may be usefully applied to cases involving nature and seasonal uncertainties.

Mixed Checking Rule

In multi-person strategic interaction, in which opponents may be engaged in checking, random checking rules may be useful. This may be accomplished by allowing $C(\omega^t)$ to be any fraction (not just 0 or 1), indicating the probability of checking after the history $\omega^t$.

The Scope of Bayesian Forecasting

Our model deals with Bayesian forecasting in the sense that the forecasted probability for the next state being s, after observing the history $\omega^t$, is the conditional probability $\tilde{\mu}(s|\omega^t)$ of some fixed probability distribution $\tilde{\mu}$. As may be seen from our definition of a forecasting rule, the scope of such forecasting is substantially broader. It covers all forecasts with this property: after every history, the forecasted values assigned to the next period possible states sum to one. For a forecast with this property, the usual construction of probabilities over finite histories (cylinder sets) leads to consistent probabilities, which may be extended uniquely in the standard way to a unique $\tilde{\mu}$ with the above mentioned properties.

Consider, for example, the 2-person fictitious-play model of a repeated game. For simplicity, assume that player I has two feasible period actions, denoted by a and b, and consider the probabilities that player II assigns to Player I's next period action. Fitting the fictitious-play forecasting rule into our terminology, a history $\omega^t$ is a t-tuple $(\omega_1,\omega_2,\ldots,\omega_t)$ with each $\omega_j$, being a or b, denoting Player I's action at period j. Player II's assessments that, in the next period, after history $\omega^t$, Player I will play the actions a and b, are computed to be their empirical frequencies in the history $\omega^t$, $(1/t)\sum_{j=1}^{t} I(\omega_j = a)$ and $(1/t)\sum_{j=1}^{t} I(\omega_j = b)$. (One must assign arbitrary starting probabilities, because the above expressions are not defined for the empty history.) Since these two numbers sum to one, we may conclude that Player II's forecast is actually Bayesian, according to some distribution $\tilde{\mu}$. ($\tilde{\mu}$ is described by a distribution of an urn process.) Thus, fictitious-play players are actually Bayesian (see Lehrer (1996) for an explicit construction).

Calibrating Versus Being Calibrated

Calibrating a forecast--i.e., bringing it to a calibrated state--may be achieved in many different ways. The learning papers cited herein offer sufficient conditions for a subjective Bayesian forecast to merge. Thus, in view of the equivalence demonstrated in this paper, we may obtain methods for calibrating through Bayesian updating.

However, other calibrating methods are used in macroeconomic models, such as those of Kydland and Prescott (1988). There one looks at large sets of past observed economic data and tries to adjust the parameters of an economic model to obtain predictions that fit these data. The hope is that if a good fit to past data is achieved, the same model will predict future outcomes well, and thus they will be calibrated. While the method of calibrating may be different, the tests of being calibrated will be similar to the ones described in this paper.

Coarser Calibration

In the model presented here the forecaster observes all past states and assigns probabilities to all possible future states. But in large economic models, such as the macroeconomic models just mentioned, this is too demanding for either side. The forecaster may only observe some events, rather than all past states, and may attempt to assign probabilities just to some future events, rather than to all states. For example, the forecaster may only forecast probabilities to several levels of future unemployment, and may only observe some restricted dimensions of the economy. These situations seem to require us to develop coarser notions of calibration. In particular, both merging and calibration under partial monitoring should be studied. A different approach could be simple restrictions of $\mu$ and $\tilde{\mu}$ to sub $\sigma$-fields, while maintaining the current notions of merging and calibration.

Calibration, Merging, and Decision Making

In economics and game theory, utility-maximizing agents use forecasts to make decisions. It is important to note that, while this paper compares forecasts by their level of calibration, a utility-maximizing agent may have other criteria.

For example, consider again the deterministically alternating 0,1 pattern of Example 1, and the naively calibrated constant (0.5,0.5) forecast $\tilde{\mu}$. Let $\tilde{\mu}_1$ be another forecast which assigns the states (0,1) the probabilities (0.9,0.1) prior to even periods (i.e., before 0 is truly realized), and the probabilities (0.05,0.95) prior to odd periods (i.e., before 1 is realized). Even though $\tilde{\mu}_1$ may be less calibrated than $\tilde{\mu}$, it may be more useful for decision making.