Comparative Testing of Experts*

Nabil I. Al-Najjar†
and
Jonathan Weinstein‡

First draft: November 2006
This version: August 2007

Abstract

We show that a simple "reputation-style" test can always identify which of two experts is informed about the true distribution. The test presumes no prior knowledge of the true distribution, achieves any desired degree of precision in some fixed finite time, and does not use "counterfactual" predictions. Our analysis capitalizes on a result due to Fudenberg and Levine (1992) on the rate of convergence of supermartingales.

We use our setup to shed some light on the apparent paradox that a strategically motivated expert can ignorantly pass any test. We point out that this paradox arises because in the single-expert setting, any mixed strategy for Nature over distributions is reducible to a pure strategy. This eliminates any meaningful sense in which Nature can randomize. Comparative testing reverses the impossibility result because the presence of an expert who knows the realized distribution eliminates the reducibility of Nature's compound lotteries.

* We are grateful to Yossi Feinberg, Drew Fudenberg, Ehud Lehrer, Wojciech Olszewski, Phil Reny, Alvaro Sandroni, Rann Smorodinsky, Muhamet Yildiz for detailed comments that substantially improved the paper. We also thank Nenad Kos for his careful proofreading.

† Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston IL 60208.
e-mail:al-najjar@northwestern.edu.
Research page : http://www.kellogg.northwestern.edu/faculty/alnajjar/htm/index.htm

‡ Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston IL 60208.
e-mail: j-weinstein@kellogg.northwestern.edu
Research page : http://www20.kellogg.northwestern.edu/facdir/facpage.asp?sid=1299

Contents

"O False and treacherous Probability,
Enemy of truth, and friend of wickednesse;
With whose bleare eyes Opinion learnes to see,
Truth's feeble party here, and barrennesse."

Keynes[1]

1 Introduction

A recent literature emerged studying whether an expert's claim to knowledge can be empirically tested. Specifically, assume that there is an unknown underlying probability distribution P generating a sequence of observations in some finite set. For example, observations may be weather conditions, stock prices, or GDP levels, while P is the true stochastic process governing changes in these variables. In each period, the expert makes a probabilistic forecast that he claims is based on his knowledge of the true process P. Can this claim be tested?

The seminal paper in this literature is that of Foster and Vohra (1998). They showed that a particular class of tests, known as calibration tests, can be passed by a strategic but totally ignorant expert.[2] Such an expert can pass a calibration test on *any* sample path without any knowledge of the underlying process. A calibration test, therefore, cannot distinguish between an informed expert who knows P and an ignorant expert. Fudenberg and Levine (1999) provided a simpler proof of this result, Lehrer (2001) generalized it to passing many calibration rules simultaneously, as do Sandroni, Smorodinsky, and Vohra (2003). Kalai, Lehrer, and Smorodinsky (1999) establish various connections to learning in games.

In a striking result, Sandroni (2003) proved the following impossibility result in a finite horizon setting: Any test that passes an informed expert can be ignorantly passed by a strategic expert on any sample path. The remarkable feature of this result is that it is not limited to any special class

[1] A Treatise on Probability, 1921.

[2] A calibration test compares the actual frequency of outcomes with the corresponding frequencies in the expert's forecast in each set of periods where the forecasts are similar. See, for example, Sandroni (2003, Sec. 3) for precise statement.

of tests, and it requires only that an expert who knows the truth can pass the test.

This disturbing result motivated a number of authors to consider models that can circumvent its conclusions. Dekel and Feinberg (2006) consider infinite horizon problems and show that there are tests that reject an ignorant expert in finite (but unbounded) time. Their positive results, however, require the use of the continuum hypothesis which is not part of standard set theory. Olszewski and Sandroni (2006) refine these findings by, among many other results, dispensing with the use of the continuum hypothesis. The tests used in these positive results do not validate a true expert in finite time. Olszewski and Sandroni (2007) prove a powerful new impossibility result showing that any test that does not condition on counterfactuals (*i.e.,* forecasts at unrealized future histories) can be ignorantly passed.

In this paper we show that these impossibility results do not extend to tests that compare two (or more) experts.[3] Our main results are stated for finite horizon testing because this is where the impossibility results are strongest and conceptually clearer.[4]

The finite horizon case therefore provides the sharpest contrast between comparative and single-expert testing. In Section 5 we show that our main results extend in sharper form to the infinite horizon case.

Our first theorem shows that in a setting with two experts there is a simple reputation-style test with the following property:[5] If one expert knows the true process P and the other is uninformed, then if the two experts make sufficiently different forecasts in sufficiently many periods, the test will pick the informed expert with high probability. The test does not rely on counterfactuals of any kind: no information about the experts' forecasts

[3] In independent work, Feinberg and Stewart (2006) also study testing multiple experts. Their work is discussed in detail in Section 5.

[4] By "finite horizon" we mean a length of time bounded independent of the true distribution or predictions made. The term "finite horizon test" is sometimes used in a different sense in the literature on testing one expert, namely as referring to tests that reject an uninformed expert in finite but not necessarily bounded amount of time. Olszewski and Sandroni (2006) show that for such tests, rejection can be delayed for as long as one wishes, limiting their applicability in practice.

[5] For expository clarity, we shall ignore quantifiers on probabilities and degrees of approximation in the introduction.

at unrealized histories is used. The theorem uses a remarkable property of the rate of convergence of martingales, discovered by Fudenberg and Levine (1992).

Case (2) of the conclusion above cannot be eliminated entirely, since an uninformed expert who randomizes will pick forecasts that are close to the truth with positive probability. The intuition, of course, is that this is an unlikely event. To make this precise, we note that the comparative test defines an incomplete-information constant-sum game between the two experts. Theorem 2 shows that the value of this constant-sum game to the uninformed player is low if the informed player is even slightly better informed, when the horizon is long enough.

Although we emphasize the finite horizon to highlight the contrast with impossibility results, most of the literature concerns the infinite horizon setting. This is indeed the case with the older calibration literature pioneered by Foster and Vohra (1998), as well as the more recent literature on general tests like Dekel and Feinberg (2006), Olszewski and Sandroni ((2006)and (2007)) and Feinberg and Stewart (2006). In Section 5 we consider the infinite horizon case and show that our main result on comparative testing extends in a stronger form.

It is important to assess the role of our assumption that there is an informed expert. In Section 4.5 we note that this assumption can be relaxed to require only that one expert has better information than the other. But this assumption cannot be dispensed with entirely: In Theorem 7 we adapt the proof of the impossibility result for the single-expert case to show that, in a finite-horizon setting, there is no non-manipulable test that can tell when there is at least one informed expert. This shows that the known limitations on single-expert testing carry over to some extent to the multiple-expert setting.

Although our primary emphasis is on comparative testing, our analysis makes a slightly more general point by shedding light on the source of the impossibility results. Roughly, we argue that the impossibility results are consequences of the facts that: (1) a stochastic process P typically has many equivalent representations, and (2) these representations are observationally indistinguishable. In the the single-expert setting, this observational

equivalence effectively impoverishes Nature's strategy sets, making it possible for a strategic expert to win. These observations provide, we believe, a unified way to understand when impossibility results are likely to obtain. For instance, impossibility results are incompatible with tests that reward information, with repeated observations of the stochastic process, or with comparison across experts, as we do here. Each of these variants works by either fully or partially restoring the richness of Nature's strategy set. Section 6 elaborates on these points.

2 Model

Fix a finite set A representing outcomes in any given period. For any set let $\Delta(\cdot)$ denote the set of probability distributions on that set.

There are n periods, $t = 1, \ldots, n$. The set of complete histories is $H^n = \big[A, \Delta(A), \Delta(A)\big]^n$, with the interpretation that the tth element $\big(a(t), \alpha_0(t), \alpha_1(t)\big)$ of a history h consists of an outcome $a(t)$, and the probabilistic forecasts $\alpha_i(t)$ of experts $i = 0, 1$ for that period.[6] Define the null history h^0 to be the empty set. A partial history of length t, denoted h^t, is any element of $\big[A, \Delta(A), \Delta(A)\big]^t \equiv H^t$.

A *time t forecasting strategy* is any $t-1$-measurable function $f^t : H^{t-1} \to \Delta(A)$, interpreted as a probabilistic forecast of the time t outcome contingent on a partial history h^{t-1}. A *forecasting strategy* $f \equiv \{f^t\}_{t=1}^n$ is a sequence of time t forecasting strategies. Two forecasts $f_i^t(h^{t-1}), i = 0, 1$, are ϵ-close if $|f_0^t(h^{t-1})(a) - f_1^t(h^{t-1})(a)| < \epsilon$ for every outcome a.

There is a true stochastic process P on A^n that generates outcomes. Let f_P be any forecasting strategy that coincides with the one-period-ahead conditionals of P at partial histories that occur with P-positive probability.

We shall think of the set of all forecasting strategies, denoted F^n, as the set of pure strategies available to an expert. Mixed strategies are probability distributions $\varphi \in \Delta(F^n)$ on the set of pure strategies.[7]

[6] To minimize repetition, from this point on, all product spaces are endowed with the product topology and the Borel σ-algebra.

[7] All probabilities on a product space are assumed to be countably additive and defined on the Borel σ-algebra generated by the product topology. Spaces of probability measures are endowed with the weak topology.

Notational Conventions. A superscript t will denote either the t-fold product of a set (as in A^t), an element of such product (*e.g.,* the vector a^t), or a function measurable with respect to the first t components of a history (*e.g.,* a time t forecast f^t or a test T^t).

An *n-period comparative test* is any measurable function[8]

$$T^n : A^n \times F^n \times F^n \to \{0, 0.5, 1\}$$

such that for every $f, f' \in F^n$ and a^n,

$$T^n(a^n, f, f') = 1 - T^n(a^n, f', f).$$

We interpret $T^n(h^n) = i$ with $i = 0, 1$, to mean that the test picks expert i after observing the history of forecasts and Nature's realizations for the past n periods. We include the value 0.5 to indicate that the test is inconclusive, in which case both experts pass.

Note the following:

- The test does not presume any structure on the underlying law;
- Each expert can condition not only on his own past forecasts and past outcomes, but also on the past forecasts of the other expert;
- The test is symmetric, in the sense that which expert is chosen by the test does not depend on the expert's label.

The test we construct below will have an additional property:

- The test does not condition on counterfactuals: given two pairs of forecasts f_0, f_1 and g_0, g_1 and a history h^n such that $f_i(h^{t-1}) = g_i(h^{t-1}), i = 0, 1$ for each t, then $T^n(a^n, f_0, f_1) = T^n(a^n, g_0, g_1)$. That is, what the experts would have forecasted at unrealized histories is not taken into account.

[8] Here, measurability is with respect to σ-algebra generated by the Borel sets on the product space H^n.

3 A Comparative Test of Experts

An expert is *truthful* if he forecasts outcomes using the true distribution P. Formally, his strategy is the deterministic forecast f_P.[9] In Theorem 7 we will show that no test can determine whether or not at least one of the two experts is truthful. Therefore the appropriate goal is a comparative test that picks a truthful expert if there is indeed one.

We introduce for each n a particular comparative test T^n as follows. Let $L_0(h^0) = 1$ and

$$L_t(h^t) = \frac{f_1^t(h^{t-1})(a(t))}{f_0^t(h^{t-1})(a(t))} L_{t-1}(h^{t-1}),^{10} \tag{1}$$

where h^t is the initial t segment of a complete history h^n, and $a(t)$ is the outcome at time t according to the history h^n. Given a history h^n Expert 1 is chosen if $L_n(h^n) > 1$, Expert 0 is chosen if $L_n(h^n) < 1$, and the test returns 0.5 (*i.e.,* it is inconclusive) if $L_n(h^n) = 1$.[11]

Theorem 1 *If expert i is truthful, then for every $\epsilon > 0$, there is an integer K such that for all integers n, distributions P, and mixed forecasting strategies φ_j $(j \neq i)$, there is $P \times \varphi_j$-probability at least $1 - \epsilon$ that either*

(a) *T^n picks expert i; or*

(b) *The two experts' forecasts are ϵ-close in all but K periods.*

Case (a) is, in a sense, the desired outcome of the test. Case (b) reflects the possibility that uninformed forecaster may get lucky and approximately guess the true law P. Note that the theorem has no bite when n is smaller than K, because case (b) will trivially obtain. The crucial point is that K is independent of the true distribution and the forecasters' strategies, so by setting n large enough case (b) says that the uninformed forecaster must have an excellent guess about the true law. Theorem 2 will support the conclusion that case (b) is "unlikely" when n is large relative to K.

[9] An expert who knows the truth may have a strategy that does better than reporting the truth; if so, this only strengthens the conclusion of Theorem 1.

[10] If the denominator is 0 in some period t, we set $L_{t'} = \infty$ for all $t' \geq t$.

[11] Using the numerical value 0.5 to denote an inconclusive outcome is convenient because it makes the Bayesian game introduced in Section 4.1 a constant sum game.

The argument relies on a result by Fudenberg and Levine (1992) establishing a uniform rate of convergence for supermartingales. See the Appendix for a formal statement of their Theorem A.1. We find it necessary to replicate part of the argument used in their main reputation result (Theorem 4.1) rather than simply citing the result because in the reputation context, the likelihood ratio, and the implied belief about types, are intermediate steps. All that matters there is the forecast of the period t short-run player about the behavior of the long-run player in period t. For us, on the other hand, the likelihood ratio is the primary object.

Proof: Without loss of generality, assume that expert 0 is truthful. It is a standard observation that the stochastic process $\{L_t\}$ is a supermartingale under the distribution induced by the strategy of Expert 0 (Lemma 4.1 in Fudenberg and Levine (1992)). As in Fudenberg and Levine, define an increasing sequence of stopping times $\{\tau_k\}_{k=0}^{\infty}$ relative to $\{L_t\}$ and ϵ inductively as follows. First, set $\tau_0 = 0$ and $\tau_k(h^\infty) = \infty$ whenever $\tau_{k-1}(h^\infty) = \infty$. If $\tau_{k-1}(h^\infty) < \infty$, let $\tau_k(h^\infty)$ be the smallest integer $t > \tau_{k-1}(h^\infty)$ such that either:

1. $P\left\{h^\infty : \left|\frac{L_t}{L_{t-1}} - 1\right| > \frac{\epsilon}{\#A} \mid h^{t-1}\right\} > \frac{\epsilon}{\#A}$; or
2. $\frac{L_t}{L_{\tau_{k-1}}} - 1 \geq \frac{\epsilon}{2\#A}$.

If there is no such t, set $\tau_k(h^\infty) = \infty$. Define the *faster process* $\{\tilde{L}_k\}$ by $\tilde{L}_k = L_{\tau_k}$ if $\tau_k < \infty$ and $\tilde{L}_k = 0$ otherwise. From standard results, the stochastic process $\{\tilde{L}_k\}$ is a supermartingale with an associated filtration whose events are generically denoted by $\tilde{h}^t$. By their Lemma 4.2, the faster process $\{\tilde{L}_k\}$ omits only observations where $|f_0^t(h^{t-1})(a) - f_1^t(h^{t-1})(a)| < \epsilon$ for all a in A.

Lemma 4.3 in Fudenberg and Levine applies, showing that $\{\tilde{L}_t\}$ is an *active supermartingale* with activity $\frac{\epsilon}{2\#A}$. We refer the reader to the appendix for formal definitions. By their Theorem A.1, for any $\epsilon > 0$ and $\#A$ there is an integer K such that for any active supermartingale $\{\tilde{L}_t\}$ with activity $\frac{\epsilon}{2\#A}$

$$P\left[\sup_{k>K} \tilde{L}_k < 1\right] > 1 - \epsilon.$$

The key point is that K depends only on ϵ and $\#A$ and not on the true stochastic process P or the forecasting strategy f_1.

Assume that Expert 1 uses a deterministic strategy. Under the assumption that Expert 0 is truthful, on a set of histories of probability $1-\epsilon$, either $|f_0^t(h^{t-1})(a) - f_1^t(h^{t-1})(a)| < \epsilon$ for all a in all but at most K periods, or $L_n < 1$.

If Expert 1 uses a mixed strategy φ, the same conclusion still follows via an application of Fubini's theorem using the facts that: (1) T^n is jointly measurable; (2) K is uniform over all forecasting strategies; and (3) P and φ_1 are independent. ■

Notice that when case (b) of Theorem 1 holds, we do not exclude the possibility that the truthful expert is rejected with probability greater than 0.5.[12] Indeed, let $n = 1, A = \{H, T\}$, and $P(H) = 0.8$. Then an expert who announces $P(H) = 0.9$ will defeat a truthful expert whenever the outcome is H, i.e. with probability 0.8. Since case (b) can be recognized by a tester without any knowledge of the truth, this issue can be resolved by making the conclusions of our test more conservative in the following natural way: for a given ϵ, modify the test to return outcome 0.5 (inconclusive) whenever the condition in case (b) holds. It is immediate that this modification does not affect the truth of Theorem 1, and satisfies the added condition that a truthful expert will be rejected conclusively with probability at most ϵ. The inconclusive verdict signifies an insufficient difference between the two experts for a statistically significant comparison at significance level ϵ.

4 The Scope of Strategic Manipulations

Theorem 1 establishes statistical properties of a simple "reputation-style" test, taking the experts' forecasts as given. That theorem does not account for experts' strategic behavior and leaves open the possibility that an uninformed expert might make a lucky guess that lands him close to the true P. This section addresses these issues.

[12] We thank Yossi Feinberg for pointing out this possibility.

4.1 A Bayesian Game

Consider the following family of incomplete-information constant-sum games between Expert 0 and Expert 1, parametrized by $n = 1, 2 \ldots$ and $\mu \in \Delta(\Delta(A^n))$:

- Nature chooses an element $P \in \Delta(A^n)$ according to a probability distribution μ;
- Expert 0 is informed of P, while Expert 1 only knows μ;
- The two players simultaneously choose forecasting strategies $f_0, f_1 \in F^n$;
- Nature then chooses a^n according to P;
- The payoff of Expert 1 is

$$T^n\big(a^n, f_0, f_1\big),$$

  where T^n is the test constructed in Theorem 1.

- The payoff of Expert 0 is $1 - T^n\big(a^n, f_0, f_1\big)$.

Payoffs are extended to mixed strategies by expected utility:

$$z(\mu, \varphi) \equiv \int_{\Delta(A^n)} \int_{F^n} \left[\int_{A^n} T^n\big(a^n, f_0, f_1\big)\, dP(a^n) \right] d\varphi(f_1)\, d\mu(P). \qquad (2)$$

4.2 The Value of the Game to the Uninformed Expert

The value of this incomplete-information constant-sum game to the uninformed player depends on how diffuse μ is. For example, if μ puts unit mass on a single $P \in \Delta(A^n)$, then the "uninformed" player knows just as much as the informed one, and so he can guarantee himself a value of 0.5. On the other hand, Theorem 1 tells us that the uninformed player can win "the reputation game" only when he succeeds in matching the true distribution in all but K periods. Our next theorem says that if μ is even slightly diffuse then his value is low when the horizon is long enough.

In the sequel, whenever convenient, we identify $\mu \in \Delta(\Delta(A^n))$ with its one-step-ahead conditionals, denoted $\mu^t(\cdot|\alpha^{t-1}) \in \Delta(\Delta(A))$. Define $\mathcal{M}(\epsilon, \delta, L) \subset \Delta(\Delta(A^n))$ to consist of all μ such that there are at least L periods $1 \leq t \leq n$ such that for μ-a.e. h^n

$$\max_{p \in \Delta(A)} \mu^t(B_\epsilon(p)|\alpha^{t-1}) < 1 - \delta.^{13} \tag{3}$$

This condition, which states that in each of at least L periods μ does not concentrate its mass in some small ball, becomes less restrictive as n becomes large.

Theorem 2 *For every ϵ and $\delta > 0$ there is an integer L such that for every $\mu \in \mathcal{M}(\epsilon, \delta, L)$ the value of the game to Expert 1 is less than ϵ.*[14]

Proof: Assume that the informed expert is required to report the truth. If he were to play strategically, the game would only become less favorable to the uninformed expert.

Let $K = K(\epsilon/2)$ be the integer obtained in Theorem 1. Let $L = L(\epsilon, \delta)$ be the smallest integer so that the binomial distribution with L trials and probability δ assigns probability at most $\frac{\epsilon}{2}$ to $\{0, \ldots, K\}$.

Fix any $\mu \in \mathcal{M}(\epsilon, \delta, L)$. It suffices to show that any fixed forecasting strategy for Expert 1 has winning probability less than ϵ. In each of the L periods described in 3, his probability of being $\frac{\epsilon}{2}$-close to the truth is at most $1-\delta$. The definition of L then guarantees that his probability of being $\frac{\epsilon}{2}$-close to the truth in all but K periods is at most $\frac{\epsilon}{2}$.

Theorem 1 tells us that when the above case does not obtain, Expert 1's probability of winning is at most $\frac{\epsilon}{2}$. We conclude that his overall winning probability is at most ϵ. ■

[13] The notation $B_\epsilon(p)$ denotes the ϵ ball around p.

[14] Oakes (1985) provided a simple argument that a Bayesian who reports his true beliefs cannot pass a calibration test on all paths. Note that, although the uninformed player in our setting has Bayesian beliefs, he is not constrained to report them truthfully. We thank a referee for bringing this result to our attention.

4.3 The Non-manipulability of Comparative Tests

Informally, the next corollary is an "anti-impossibility" result: It says that if one expert knows Nature's distribution, an uninformed strategic expert cannot guarantee success simultaneously against all distributions. That is, for any mixed strategy over forecasts, Nature has a distribution $P \in \Delta(A^n)$ such that the uninformed expert passes the test with probability at most ϵ.

Corollary 3 *For every ϵ and $\delta > 0$ there is an integer L such that for every $\mu \in \mathcal{M}(\epsilon, \delta, L)$ and every φ_1 there is $P \in \operatorname{supp} \mu$ such that*

$$z(P, \varphi_1) < \epsilon.$$

Proof: From Theorem 2 we have

$$z(\mu, \varphi_1) < \epsilon.$$

Then there must be an element in $P \in \operatorname{supp} \mu$ such that the conclusion of the theorem holds. ■

4.4 What does it mean to be Uninformed?

Consider three environments that would look identical to an uninformed expert in the absence of an informed one:

- $\hat{\mu}$ is characterized by $\hat{\mu}^t(\cdot|\alpha^{t-1})$ being the uniform distribution, independently across partial histories, on the vertices of $\Delta(A)$;
- $\bar{\mu}$ is characterized by $\bar{\mu}^t(\cdot|\alpha^{t-1})$ being the uniform distribution, independently across partial histories, over a small ball around the distribution $\bar{p}$ that assigns equal probability to all outcomes;
- $\tilde{\mu}$ is defined similarly, except that $\tilde{\mu}^t(\cdot|\alpha^{t-1})$ puts unit mass on $\bar{p}$.

Fix a sufficiently large n so that $\bar{\mu}$ and $\hat{\mu}$ defined above both belong to $\mathcal{M}(\epsilon, \delta, L)$ for some $\epsilon, \delta > 0$ and L as in Theorem 2.

The first point to make is that our assumption that the informed player knows the true distribution P is not as strong as it might first appear. Under

$\hat{\mu}$ the informed player knows the deterministic path of outcomes, and so he knows as much as there is to be known. By comparison, the informed player under $\bar{\mu}$ or $\tilde{\mu}$ knows much less, yet we still refer to him as "informed."

Our second point is that in stochastic environments the relevant measure of being (un)informed is relative. Under $\tilde{\mu}$ both players are uninformed, and so they achieve equal value of 0.5. Under $\bar{\mu}$ the informed player is only slightly more informed, yet this is enough to tilt the game in his favor.

In summary, the uninformed experts in these three environments have identical beliefs over realized events and so in any single-expert test they would necessarily perform equally well. On the other hand, their performance in comparative tests vary widely. These differences in performance in a comparative test stem from how much they know *relative* to their opponents. This supports our view that any identifiable notion of truth is inherently relative: In recognizing a stochastic truth we cannot do better than to define it as the belief of the most knowledgeable expert.

4.5 Exact vs. Better Knowledge of the Truth

We have focused exclusively on the case in which one expert knows the true probability law and the other does not. It is natural to wonder whether our results extend to the case that neither knows the whole truth, but one knows more than the other. In fact, with an appropriate definition of "knowing more," it is fairly easy to show that these cases are indistinguishable. Assume the possible states of nature are described by a space Ω, and to each $\omega \in \Omega$ there corresponds a distribution P_ω on A^n. Now suppose the experts have a common prior Q on Ω, each expert's knowledge of the state is described by a partition Π_i, and the partition of expert 0 is strictly finer. In this model, expert 0 does not know the precise state ω and thus does not know the true probability distribution P_ω, but the model is completely equivalent to the following "quotient model" in which he does: let the new state space be Π_0, and the distribution corresponding to each state (partition element) be simply the distribution on A^n conditional on that partition element in the original model. To the tester, this model is indistinguishable from the original since no one can observe ω more accurately than expert 0– in fact the changes are purely formal, but in the new model expert 0 knows

the true probabilities.

5 Infinite Horizon

So far we have confined ourselves to the finite horizon setting because it provides the sharpest contrast between the one- and two-experts cases. Most of our results in fact extend to the infinite horizon in stronger form.

The comparative test can be extended by first defining the process $L_t(h^t)$ exactly as in 1. In defining the test we need to account for the possibility that L_t might not converge. Thus, the test chooses Expert 0 if $\lim_{n\to\infty} L_n(h^n) < 1$, Expert 1 is chosen if $\lim_{n\to\infty} L_n(h^n) > 1$, and chooses an expert at random if either $\lim_{n\to\infty} L_n(h^n) = 1$ or this sequence fails to converge.[15] The constant K derived in Theorem 1 is independent of the horizon.[16]

In the infinite horizon case we obtain the sharper result that either an informed expert is picked, or the two experts asymptotically make identical forecasts:

Theorem 4 *If expert i is truthful, then for any distribution P and mixed forecasting strategy φ_j ($j \neq i$), there is $P \times \varphi_j$-probability 1 that either*

(a) *T picks expert i; or*

(b) $\lim_{t\to\infty} |f_0^t(h^{t-1}) - f_1^t(h^{t-1})| = 0$.

Proof: Assume without loss of generality that Expert 0 is truthful, and fix arbitrary P and f_1. Write $\epsilon_n \equiv \frac{1}{2^n}$ and repeatedly apply Theorem 1 to obtain a sequence of integers $\{K_n\}$ such that each event

$$A_n \equiv \left\{h : \lim_{t\to\infty} L_t > 1 \;\&\; \#\{t : |f_0^t(h^{t-1}) - f_1^t(h^{t-1})| > \epsilon_n\} > K_n\right\}$$

[15] By the martingale convergence theorem the last case occurs with zero probability if there is an informed expert.

[16] As it is in Fudenberg and Levine (1992) active supermartingale result.

has probability less than ϵ_n.[17] , [18] Since $\sum_n P(A_n) < \infty$, by the Borel-Cantelli Lemma we have:

$$P\{h \in A_n \text{ i.o.}\} = 0.$$

Thus, with P-probability 1 along each path h, either Expert 0 wins or $|f_0^t(h^{t-1}) - f_1^t(h^{t-1})| \leq \epsilon_n$ for all but K_n periods. In the latter case $|f_0^t(h^{t-1}) - f_1^t(h^{t-1})| \to 0$. ■

Our final result shows that Theorem 2 extends to the infinite horizon in a sharper form. The Bayesian game in Section 4.1 extends to the infinite horizon setting.[19] As in Section 4.1 we identify $\mu \in \Delta(\Delta(A^\infty))$ with its one-step-ahead conditionals, denoted $\mu^t(\cdot|\alpha^{t-1}) \in \Delta(\Delta(A))$. Define $\mathcal{M}(\epsilon, \delta) \subset \Delta(\Delta(A^\infty))$ to consist of all μ such that for μ-a.e. infinite history h^∞, for infinitely many periods,

$$\max_{p \in \Delta(A)} \mu^t(B_\epsilon(p)|\alpha^{t-1}) < 1 - \delta. \tag{4}$$

Theorem 5 *For every $\epsilon,\ \delta > 0$ and $\mu \in \mathcal{M}(\epsilon, \delta)$ the value of the game to Expert 1 is zero.*

Proof: The proof closely follows that of Theorem 2 so assume, as in that proof, that the informed expert reports the truth.

It suffices to show that the payoff of the strategic expert is 0 for each of his pure strategies f_1. For any pair of integers K and L, we have

$$\mu\Big\{f_0 : \#\{t : |f_0^t(h^{t-1}) - f_1^t(h^{t-1})| > \epsilon\} \leq K\Big\} < B(K, L, \delta)$$

[17] We note that it is evident from the proof of Theorem 1 that finiteness of the horizon is superfluous to that theorem; the argument used in its proof can be cast in an infinite horizon to draw the conclusion that for every ϵ there is K such that for every integer n with P-probability at least $1 - \epsilon$ either

(a) $L_n < 1$; or

(b) The two experts' forecasts are ϵ-close in all but K periods.

[18] By appealing to Footnote 15, we ignore the possibility that L_t might not converge.

[19] The details are standard. The sets of infinite realizations A^∞ and histories H^∞ are given the product topologies. Probabilities on these spaces are defined on the Borel σ-algebras on these spaces. Mixed strategies are defined on the Borel σ-algebra generated by the weak topology on $\Delta(A^\infty)$.

where $B(K, L, \delta)$ denotes the binomial probability of no more than K successes in L trials when the probability of success is δ. Taking L to infinity (holding K fixed), the RHS goes to 0. Therefore the LHS is equal to zero for every K.

$$\mu\{f_0 : |f_0^t(h^{t-1}) - f_1^t(h^{t-1})| > \epsilon \ \ i.o.\} = 1$$

so Case b in Theorem 4 holds with probability 0. The payoff of the strategic expert is therefore 0. ■

We now discuss the recent work of Feinberg and Stewart (2006) who take an alternative approach to testing multiple forecasters. Their test, called "cross-calibration," extends the standard calibration test by requiring that a potential expert give frequencies that are correct in the infinite limit not just conditional on his own forecast, but conditional on any combination of his and the other player's forecasts. In addition to the choice of calibration versus reputation-style testing, their methodology differs from ours in that they work with an infinite horizon, the case we consider in this section, which has the advantage of allowing zero probability of random error. We find it instructive to consider how much error can be reduced in a finite horizon. They also have a different framework for evaluation of the effectiveness of a test, namely the topological notion of category (also used by Dekel and Feinberg (2006) and Olszewski and Sandroni (2006)). Their central result shows that when a false expert is cross-calibrated against a true expert, for any strategy he might use he will pass with positive probability only on a category 1 set of true distributions. The category approach has the advantage of not requiring the (perhaps arbitrary) specification of a distribution over distributions to represent the false expert's uncertainty about the true probabilities. Nevertheless, since a classical decision-maker wants to evaluate the overall (subjective) probability with which he passes rather than the category of his passing set, we find it instructive to consider possible second-order distributions for the false expert and evaluate which ones are concentrated enough to allow him to pass with high probability.

6 Discussion

For expositional clarity, we shall refer to the forecaster's pure strategies as measures $Q \in \Delta(A^n)$, so his set of mixed strategies is $\Delta(\Delta(A^n))$, exactly the same as Nature's.

6.1 Key intuition underlying impossibility results

We begin with an informal review of the typical minimax argument used to prove impossibility. Our prototype is Sandroni (2003)'s disarmingly elegant argument, which we informally outline.

In the single-expert setting a test is a function of the form:

$$T_s^n : A^n \times \Delta(A^n) \to \{0, 1\}$$

with the interpretation that the test decides whether or not to pass the expert based on the sequence of outcomes a^n and the expert's forecast $Q \in \Delta(A^n)$. A strategic expert's payoff is the expected probability of passing the test:

$$z_s(P, \varphi) = \int_{A^n} \int_{\Delta(A^n)} T_s^n(a^n, Q)\, d\varphi(Q)\, dP(a^n).$$

Here, expectation is taken with respect to the expert's randomization φ over forecasts and Nature's randomization over the sequence of outcomes a^n.

An impossibility result asserts that the expert has a strategy φ that guarantees him a high payoff regardless of what Nature does. Think of the forecaster as playing a constant-sum game against Nature, in which case the Minimax Theorem asserts:

$$\max_{\varphi \in \Delta(\Delta(A^n))} \min_{P \in \Delta(A^n)} z_s(P, \varphi) = \min_{P \in \Delta(A^n)} \max_{\varphi \in \Delta(\Delta(A^n))} z_s(P, \varphi). \tag{5}$$

The impossibility theorem boils down to putting a lower bound on the maxmin value in the above expression.

This is where the crucial assumption that a test must pass the truth comes into play. Formally, a test T_s^n *passes the truth with probability* $1 - \epsilon$ if:

$$z_s(P, P) \equiv P\{T_s^n(a^n, P) = 1\} > 1 - \epsilon. \tag{6}$$

This condition ensures that the RHS of Eq. 5 is close to 1: if the expert knew that Nature has chosen P, then he has an obvious response guaranteeing a payoff of $1-\epsilon$, namely to report P. This delivers the conclusion that the maxmin value is also greater than $1-\epsilon$, *i.e.,* the strategic expert can pass the test with high probability.

To summarize, the impossibility theorem consists of two key ingredients:

- The Minimax Theorem;
- The assumption that the test must pass the truth.

We examine these in turn.

6.2 Nature's Strategies and the Minimax Theorem

In a game between an expert and Nature, mixed strategies $\mu, \varphi \in \Delta(\Delta(A^n))$ are two stage lotteries. Let $P_\mu, Q_\varphi \in \Delta(A^n)$ denote the corresponding probability measures obtained from μ and φ through the usual reduction of compound lotteries.

In the single-expert setting one may write the conclusion of the Minimax theorem as:

$$\max_{\varphi\in\Delta(\Delta(A^n))} \min_{\mu\in\Delta(\Delta(A^n))} z_s(\mu,\varphi) = \min_{\mu\in\Delta(\Delta(A^n))} \max_{\varphi\in\Delta(\Delta(A^n))} z_s(\mu,\varphi). \quad (7)$$

But Nature's randomization in this case is completely superfluous. As far as the payoffs are concerned, whether Nature uses a mixed strategy μ or its equivalent pure strategy reduction P_μ makes no difference:

$$z_s(\mu,\varphi) = z_s(P_\mu,\varphi), \quad \forall \mu,\varphi \in \Delta(\Delta(A^n)). \quad (8)$$

This is because μ and P_μ induce identical distributions on the set of outcomes A^n. As far as realized outcomes are concerned, μ and P_μ are observationally indistinguishable. For example, an outside observer (in particular, the test) can never distinguish between whether Nature is playing a 50/50 lottery on two measures P_1 and P_2 or putting unit mass on the measure $P_\mu = \frac{P_1+P_2}{2}$.

By contrast, in general, an expert's mixed strategy ν is not reducible in the same manner: choosing between the two forecasts Q_1 or Q_2 with equal probability is not payoff equivalent to the forecast $Q = \frac{Q_1+Q_2}{2}$.

The crucial consequence of this asymmetry between Nature's and the expert's randomization is that the values appearing in Eq. 5 and 7 coincide:

$$\min_{\mu\in\Delta(\Delta(A^n))} \max_{\varphi\in\Delta(\Delta(A^n))} z_s(\mu,\varphi) = \min_{P\in\Delta(A^n)} \max_{\varphi\in\Delta(\Delta(A^n))} z_s(P,\varphi).$$

This effectively impoverishes Nature's strategy sets, making it possible for a strategic expert to win.

Our results on comparative testing may be understood as a consequence of the restoration of $\Delta(\Delta(A^n))$ as Nature's strategy space. To facilitate comparison with the single-expert literature, think of a constant-sum game where Nature uses a mixed strategy μ and informs Expert 0 of its random choice $P \in \Delta(A^n)$. Unless μ is degenerate, Nature's use of a mixed strategy μ is strategically distinct from P_μ, in the sense that Eq. 8 no longer holds. To win, the strategic expert has to guess Nature's selection of a pure strategy.

At more abstract level, in both the single- and multiple-expert problems: (1) the strategic expert has a "good response" to any of Nature's probability distributions, but (2) Nature has the opportunity to randomize over these distributions. The crucial difference is that in the single-expert case, having a good response to any distribution is equivalent to having a good response to any randomization over distributions, since the two are equivalent via the reduction of compound lotteries. In the multiple-expert case this equivalence no longer holds.

6.3 Passing the Truth and the Value of Information

A striking aspect of the impossibility results is the weakness of its assumptions. Aside from structural assumptions, the only requirement is that an expert who knows the true distribution should pass with probability $1-\epsilon$. This seemingly weak and compelling requirement is more subtle and powerful than it might initially appear.

To appreciate its power, think of the hide-and-seek game where Nature "hides" the true probability law P somewhere in the convex set $\Delta(A^n)$; the expert's task is to find the hidden P. With many (in fact, infinite) locations for hiding, the hider should have the advantage in such a game. Yet the impossibility results say that the seeker (the strategic expert) has the upper hand. How can that be?

The discussion in the last subsection explains this puzzle: A randomized hiding location μ by Nature is equivalent to it choosing the deterministic expected hiding location P_μ. The expert, on the other hand, can randomize his search, negating the hider's advantage.

Where does that leave us with the assumption that a test must pass the truth? There is clearly no ambiguity in the meaning of a deterministic truth. The meaning of stochastic truth, as the quote from Keynes suggests, is much less obvious. A typical distribution P on outcomes can have infinitely many two-stage lottery representations μ (with $P_\mu = P$). Different representations correspond to meaningful and distinct information structures. But these different information structures are relevant only to the extent that there is an observer who is at least partially informed of what the truth is.

7 Concluding Remarks: *Isolated vs. Comparative Testing*

Impossibility results provide invaluable insights by uncovering the subtle consequences of their assumptions. In this sense, Sandroni (2003)'s theorem revealed how innocuous-looking properties of the testing environment make it impossible to test probabilistic theories. That any test can be passed by a strategic expert is a profoundly disturbing message to the countless areas of human activity where testing experts' knowledge is vital.

In this paper we construct tests with good properties by departing from the assumption that forecasts are tested in isolation. We also use the model of comparative testing to shed light on what makes the impossibility result possible and, thus, what it takes to avoid it.

How are experts and their theories tested in practice? We are unaware of any comprehensive study, but it is not hard to identify regularities in specific contexts. The human activity where testing theories is handled with the greatest care and rigor is, arguably, scientific knowledge.[20] There are

[20]The impossibility results seem to undermine the central methodological principle of falsifiability as a criterion for judging whether a theory is scientific or not. The impossibility results imply that given any rule of evaluating scientific theories, a strategic experts can produce a falsifiable theory Q that is very unlikely to be rejected by that rule, regardless of what the truth is. Harman and Kulkarni (2007) provide a different perspective and

numerous and well-known examples where theories are judged in terms of their performance relative to other theories rather than in isolation. Some of the greatest scientific theories were, or continue to be, maintained despite a large body of contradicting evidence. A well-known example is Newtonian gravitational theory which was upheld for decades despite various empirical anomalies. This theory was eventually replaced, but only as a consequence of a comparison with a better theory, general relativity. Perhaps less known to the reader is the steady accumulation of empirical findings inconsistent with general relativity—as well as its fundamental incompatibility with other theories in physics. Yet this theory continues to be maintained because no other theory does better.[21] Economics is full of similar examples. Expected utility theory continues to be the dominant theory in economic models despite the overwhelming evidence against it. The reason, we suspect, is the lack of a convincing alternative.

In practice, comparative testing is common and, arguably, a more prevalent method of testing theories. Weather forecasters, stock analysts, and macroeconomists can be, and often are, judged relative to each other, not according to some absolute pass/fail test. Our results provide a very simple reputation-type approach to conducting such comparative tests.

To conclude, an interpretation of the impossibility literature, combined with our positive results for comparative testing, is that the only coherent notion of "true" probabilities is relative. That is, we cannot say whether or not a theory is correct in any absolute sense, only that it is better than others.

discuss the limitations of simplistic popperian falsifiability when theories are probabilistic.

[21] For details on these examples, see Darling (2006).

A Appendix

A.1 The Active Supermartingale Theorem

Consider an abstract setting with a probability measure P on H^∞ and a filtration $\{\mathcal{H}_k\}_{k=1}^\infty$ where each $\mathcal{H}_k$ is generated by a finite partition, with generic element denoted $\tilde{h}^k$.

Definition 1 *A positive supermartingale $\{\tilde{L}_k\}$ is* active *with activity $\psi > 0$ (under P) if*

$$P\left\{h^\infty : \left|\frac{\tilde{L}_k}{\tilde{L}_{k-1}} - 1\right| > \psi \mid \tilde{h}^{k-1}\right\} > \psi$$

for almost all histories with $\tilde{L}_{k-1} > 0$.

Fudenberg and Levine (1992, Theorem A.1) show the following remarkable result:

Theorem 6 *For every $l_0 > 0, \epsilon > 0, \psi \in (0,1)$ and $0 < \bar{L} < l_0$ there is a time $K < \infty$ such that*

$$P\left\{h^\infty : \sup_{k>K} \tilde{L}_k \leq \bar{L}\right\} \geq 1 - \epsilon$$

for every active supermartingale $\{\tilde{L}_k\}$ with $\tilde{L}_0 = l_0$ and activity ψ.

The power of the theorem stems from the fact that the integer K, which depends on the parameters l_0, ϵ, ψ and $\bar{L}$, is otherwise independent of the underlying stochastic process P. Note that $\tilde{L}_k$, being a supermartingale, is weakly decreasing in expectations. The assumption that it is active says that it must substantially go *up or down* relative to $\tilde{L}_{k-1}$ with probability bounded away from zero in each period. The theorem says that if $\{\tilde{L}_k\}$ is an active supermartingale, then there is a fixed time K by which, with high probability, $\tilde{L}_k$ drops below $\bar{L}$ and *remains* below $\bar{L}$ for all future periods.

The result has important applications in the reputation literature and is also related to the concept of weak merging, introduced by Kalai and Lehrer (1994). Sorin (1999) introduces a framework that integrates the reputation and merging literatures.

In the context of testing, we consider two strategies, one for each expert. Although the testing context is not inherently Bayesian, the tester is free to design a test with Bayesian features, where the forecasting strategies correspond to 'types' and 'beliefs' are updated using Bayes rule. Our comparative test chooses an expert depending on whether the posterior odds ratio is above or below 1. The active martingale result implies that there is a bound (independent of the length of the game and the true distribution) on the number of periods where the uninformed expert can be substantially wrong, such that if this bound is exceeded, the probability that $L_n > 1$ is small.

Our use of the active supermartingale result differs from the reputation model in another way. There it was necessary to show that, should beliefs over actions differ too often, L_n will fall close to zero, implying that the uninformed player would be almost certain he is facing the commitment type, whereas here we are only interested in whether L_n rises or falls marginally over the horizon of the model.

A.2 Impossibility of testing whether there is at least one informed expert

We now consider the issue of whether there is a way to determine if among the two experts at least one is informed. Formally, consider a function

$$\tau : H^n \to \{0, 1\}$$

with the interpretation that $\tau(a^n, f_0, f_1) = 1$ iff at least one expert is informed.[22]

The following theorem is an important consequence of Sandroni (2003)'s impossibility result:

Theorem 7 *Suppose that τ is such that for every P, f_0 and f_1*

$$P\{a^n : \tau(a^n, f_0, f_1) = 1\} > 1 - \epsilon \quad \textit{if either } f_0 = f_P \textit{ or } f_1 = f_P. \tag{9}$$

[22]Note that we allow the test τ to condition on the entire forecasting schemes, including forecasts at unobserved histories. This only strengthens the conclusion of Theorem 7.

Then for every mixed strategy φ_0 of Expert 0 there is a mixed strategy φ_1 of Expert 1 such that for every a^n

$$\varphi_0 \times \varphi_1\big\{(f_0, f_1) : \tau(a^n, f_0, f_1) = 1\big\} > 1 - \epsilon. \tag{10}$$

That is, if τ has the property that it returns 1 (with high probability) whenever at least one expert is informed, then each of the two experts can, for any opponent strategy, manipulate τ by forcing it to return 1 (with high probability) without any knowledge of the true process.

Proof: For any forecasting strategy f_0 of Expert 0, define the single-expert test

$$M_{f_0} : A^n \times F^n \to \{0, 1\}$$

by

$$M_{f_0}(a^n, f_1) = 1 \iff \tau(a^n, f_0, f_1) = 1.$$

By 9, the single-expert test M_{f_0} passes the truth with probability $1 - \epsilon$. From Sandroni (2003) we know that there is a mixed strategy φ_1 such that for every a^n

$$\varphi_1\{f_1 : M_{f_0}(a^n, f_1) = 1\} > 1 - \epsilon.$$

This establishes 10 for pure φ_0.

For a general φ_0 Expert 1 is facing a lottery over deterministic tests. We show that Sandroni (2003)'s impossibility result extends to the case of stochastic tests. Formally, for each a^n and f_1 define the single-expert test

$$M_{\varphi_0}(a^n, f_1) \equiv \varphi_0\{f_0 : \tau(a^n, f_0, f_1) = 1\}.$$

The reader may interpret M_{φ_0} as either a score in a continuous valued test, or as the probability chosen by the tester to pass the expert at a^n and f_1.

Note that for any f_1

$$\begin{aligned}
\int_{A^n} M_{\varphi_0}(a^n, f_1)\, dP_{f_1} &\equiv \int_{A^n} \int_{f_0} \tau(a^n, f_0, f_1)\, d\varphi_0\, dP_{f_1} \\
&= \int_{f_0} \int_{A^n} \tau(a^n, f_0, f_1)\, dP_{f_1}\, d\varphi_0 \\
&= \int_{f_0} P_{f_1}\{a^n : \tau(a^n, f_0, f_1) = 1\}\, d\varphi_0 \;>\; 1 - \epsilon.
\end{aligned}$$

Applying the Minimax Theorem (Fan (1953)), we conclude that there is φ_1 such that for every a^n

$$\varphi_1\{f_1 : M_{\varphi_0}(a^n, f_1) = 1\} > 1 - \epsilon,$$

from which 10 directly follows. ■

Theorem 7 does *not* extend to the infinite horizon–at least not without additional restrictions. This is because a key ingredient of its proof is the impossibility result for finite-horizon testing. In the infinite-horizon case there are a number of positive results, as noted in the introduction. However, Olszewski and Sandroni (2007) prove an impossibility theorem for all infinite-horizon tests that do not use counterfactuals.

References

Darling, D. (2006): *Gravity's Arc*. Wiley, New York.

Dekel, E., and Y. Feinberg (2006): "Non-Bayesian Testing of an Expert," *Review of Economic Studies*, 73, 893–906.

Fan, K. (1953): "Minimax theorems," *Proc. Nat. Acad. Sci. U. S. A.*, 39, 42–47.

Feinberg, Y., and C. Stewart (2006): "Testing Multiple Experts," Yale and Stanford.

Foster, D., and R. Vohra (1998): "Asymptotic calibration," *Biometrika*, 85(2), 379–390.

Fudenberg, D., and D. K. Levine (1992): "Maintaining a reputation when strategies are imperfectly observed," *Review of Economic Studies*, 59(3), 561–579.

Fudenberg, D., and D. K. Levine (1999): "An Easier Way to Calibrate," *Games and Economic Behavior*, 29(1), 131–137.

Harman, G., and S. Kulkarni (2007): *Reliable Reasoning: Induction and Statistical Learning Theory*. MIT Press.

Kalai, E., and E. Lehrer (1994): "Weak and Strong Merging of Opinions," *Journal of Mathematical Economics*, 23, 73–86.

Kalai, E., E. Lehrer, and R. Smorodinsky (1999): "Calibrated Forecasting and Merging," *Games and Economic Behavior*, 29(1), 151–159.

Lehrer, E. (2001): "Any Inspection Is Manipulable," *Econometrica*, 69(5), 1333–1347.

Oakes, D. (1985): "Self-Calibrating Priors Do Not Exist," *Journal of the American Statistical Association*, 80(390), 339–339.

Olszewski, W., and A. Sandroni (2006): "Strategic Manipulation of Empirical Tests," Northwestern University.

——— (2007): "Counterfactual Predictions," Northwestern University.

Sandroni, A. (2003): "The reproducible properties of correct forecasts," *Internat. J. Game Theory*, 32(1), 151–159.

Sandroni, A., R. Smorodinsky, and R. Vohra (2003): "Calibration with Many Checking Rules," *Mathematics of Operations Research*, 28(1), 141–153.

Sorin, S. (1999): "Merging, reputation, and repeated games with incomplete information," *Games Econom. Behav.*, 29(1-2), 274–308.