

# ECONOMETRICA

JOURNAL OF THE ECONOMETRIC SOCIETY

*An International Society for the Advancement of Economic  
Theory in its Relation to Statistics and Mathematics*

<http://www.econometricsociety.org/>

*Econometrica*, Vol. 76, No. 3 (May, 2008), 541–559

## COMPARATIVE TESTING OF EXPERTS

NABIL I. AL-NAJJAR

*Kellogg School of Management, Northwestern University, Evanston, IL 60208, U.S.A.*

JONATHAN WEINSTEIN

*Kellogg School of Management, Northwestern University, Evanston, IL 60208, U.S.A.*

---

The copyright to this Article is held by the Econometric Society. It may be downloaded, printed and reproduced only for educational or research purposes, including use in course packs. No downloading or copying may be done for any commercial purpose without the explicit permission of the Econometric Society. For such commercial purposes contact the Office of the Econometric Society (contact information may be found at the website <http://www.econometricsociety.org> or in the back cover of *Econometrica*). This statement must be included on all copies of this Article that are made available electronically or in any other format.

---

## COMPARATIVE TESTING OF EXPERTS

BY NABIL I. AL-NAJJAR AND JONATHAN WEINSTEIN<sup>1</sup>

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 of 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.

KEYWORDS: Testing, reputation, probability.

*O False and treacherous Probability,  
Enemy of truth, and friend of wickedness;  
With whose bleare eyes Opinion learns to see,  
Truth’s feeble party here, and barrenness.*

Keynes, *A Treatise on Probability* (1921)

### 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  $P$  that generates 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.<sup>2</sup> Such an expert can pass a calibration test on *any* sample path without any knowledge of the underlying process.

<sup>1</sup>We are grateful to Yossi Feinberg, Drew Fudenberg, Ehud Lehrer, Wojciech Olszewski, Phil Reny, Alvaro Sandroni, Rann Smorodinsky, Muhamet Yildiz for their detailed comments. The paper substantially improved as a result of the detailed and thoughtful comments by co-editor Larry Samuelson and three anonymous referees. We also thank Nenad Kos and Jie Gong for their careful proofreading.

<sup>2</sup>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, Section 3) for precise statement.

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, while Lehrer (2001) and Sandroni, Smorodinsky, and Vohra (2003) generalized it to passing many calibration rules simultaneously. Kalai, Lehrer, and Smorodinsky (1999) established various connections to learning in games.

Sandroni (2003) proved the following striking impossibility result in a finite horizon setting: Any test that passes all informed experts 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 a special class of tests: 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) considered infinite-horizon problems and showed 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) refined 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) proved 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 reconsider these impossibility results in the context of testing multiple experts.<sup>3,4</sup> Our first theorem shows that in a finite-horizon setting with two experts there is a simple reputation-style test with the following property<sup>5</sup>: 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 at unrealized histories is used. The theorem uses a remarkable property of the rate of convergence of supermartingales that was discovered by Fudenberg and Levine (1992).

Our result cannot rule out the possibility that the test picks an uninformed expert, since such an expert may randomly select a forecast that is close to the truth. 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

<sup>3</sup>In independent work, Feinberg and Stewart (2006) also studied testing multiple experts. Their work is discussed in detail in Section 5.

<sup>4</sup>Although our main results are stated for the case of two experts, they have straightforward extensions to testing  $n$  experts by simply selecting the expert with the highest likelihood.

<sup>5</sup>For expository clarity, we shall ignore quantifiers on probabilities and degrees of approximation in the Introduction.

of this constant-sum game to the uninformed player is low if the informed player is even slightly better informed (in a sense to be made precise) and the horizon is long enough.

Our main results are stated for finite-horizon testing because this is where the impossibility results are strongest and conceptually clearest.<sup>6</sup> On the other hand, most of the literature, for example, on calibration tests, concerns the infinite-horizon setting. In Section 5 we consider the infinite-horizon case and show that our main results on comparative testing extend 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 nonmanipulable test that can tell whether there is at least one informed expert. This shows that notwithstanding our effective comparative test, the known limitations on single-expert testing still have force in 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 any stochastic process  $P$  has many equivalent representations, and these representations are observationally indistinguishable in the single-expert setting. This observational equivalence effectively impoverishes Nature's strategy sets, making it possible for a strategic expert to win. This provides a way to understand why impossibility results fail in certain circumstances, such as under repeated observations of the stochastic process or when comparing experts as in this paper. In each of these variants, the richness of Nature's strategy set is at least partially restored. Section 6 elaborates on these points.

## 2. MODEL

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

There are finitely many periods,  $t = 1, \dots, n$ . The set of complete histories is  $H^n = [A, \Delta(A), \Delta(A)]^n$ , with the interpretation that the  $t$ th element  $(a(t), \alpha_0(t), \alpha_1(t))$  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.<sup>7</sup> Define the null history

<sup>6</sup>By "finite horizon" we mean a length of time bounded independently of the true distribution or predictions made. The term "finite-horizon test" is sometimes used in a different sense in the literature, referring to tests that reject an uninformed expert in a finite but not necessarily bounded amount of time. Olszewski and Sandroni (2006) showed that for such tests, rejection can be delayed for as long as one wishes, limiting their applicability in practice.

<sup>7</sup>To minimize repetition, from this point on, all product spaces are endowed with the product topology and the Borel  $\sigma$ -algebra.

$h^0$  to be the empty set. A partial history of length  $t$ , denoted  $h^t$ , is any element of  $[A, \Delta(A), \Delta(A)]^t \equiv H^t$ .

A *time  $t$  forecasting strategy* is any  $(t - 1)$ -measurable function  $f^t: H^{t-1} \rightarrow \Delta(A)$ , interpreted as a 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  $\varepsilon$ -close if  $|f_0^t(h^{t-1})(a) - f_1^t(h^{t-1})(a)| < \varepsilon$  for every outcome  $a$ .

Any forecasting strategy  $f$  defines a unique stochastic process  $P_f$  on  $A^n$  in the obvious way. Conversely, given a stochastic process  $P$ , we 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.<sup>8</sup>

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.<sup>9</sup> We shall assume that all randomizations by Nature and the experts are independent.

*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<sup>10</sup>

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

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 first  $n$  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 probability law.

<sup>8</sup>We will follow the customary practice of identifying a stochastic process with its one-step-ahead conditionals. Note that this is not entirely innocuous in a testing context, since a test that takes a forecasting strategy  $f$  as input could, in principle, condition on forecasts at histories that have zero probability under  $P_f$ . This possibility is not relevant for the comparative test we introduce in this paper.

<sup>9</sup>All probabilities on a product space are assumed to be countably additive and defined on the Borel  $\sigma$ -algebra generated by the product topology. Spaces of probability measures are endowed with the weak topology.

<sup>10</sup>Here, measurability is with respect to  $\sigma$ -algebra generated by the Borel sets on the product space  $H^n$ .

- 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 forecast at unrealized histories is not taken into account.

### 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  $f_P$ .<sup>11</sup> A natural question is whether there exists a test that can determine if at least one expert is truthful. Theorem 7 in the Appendix shows that no such test exists. Therefore, the appropriate goal, and the focus of our paper, is a comparative test that picks a truthful expert if indeed there is one.

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

$$(1) \quad L_t(h^t) = \frac{f_1^t(h^{t-1})(a(t))}{f_0^t(h^{t-1})(a(t))} L_{t-1}(h^{t-1}),^{12}$$

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  $L_n(h^n) = 1$ .<sup>13</sup>

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

- (a)  $T^n$  picks Expert  $i$  or
- (b) the two experts’ forecasts are  $\varepsilon$ -close in all but  $K$  periods.

Case (a) is, in a sense, the desired outcome of the test. Case (b) reflects the possibility that an uninformed forecaster may get lucky and approximately guess the true law  $P$ . Note that the theorem has no bite when  $n$  is smaller than

<sup>11</sup>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.

<sup>12</sup>If the denominator is 0 in some period  $t$ , we set  $L_{t'} = \infty$  for all  $t' \geq t$ .

<sup>13</sup>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.

$K$ , 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  $K$ .

The argument relies on a result by Fudenberg and Levine (1992) that established a uniform rate of convergence for supermartingales.<sup>14</sup>

**PROOF OF THEOREM 1:** Without loss of generality, assume that Expert 0 is truthful. In the proof, it will be convenient to work with infinite histories, although the test conditions only on the first  $n$  periods, for a fixed  $n$ .

It is a standard observation that the stochastic process  $\{L_t\}$  is a supermartingale under  $P$  (Lemma 4.1 in Fudenberg and Levine (1992), henceforth FL). As in FL, define an increasing sequence of stopping times  $\{\tau_k\}_{k=0}^\infty$  relative to  $\{L_t\}$  and  $\varepsilon$  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.  $P(h^{t-1}) > 0$  and  $P\{h^\infty : |L_t/L_{t-1} - 1| > \varepsilon/\#A | h^{t-1}\} > \varepsilon/\#A$  or
2.  $L_t/L_{\tau_{k-1}} - 1 \geq \varepsilon/(2\#A)$ .

If there is no such  $t$ , set  $\tau_k(h^\infty) = \infty$ . Define the 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. By FL's Lemma 4.2,  $|f_0^t(h^{t-1})(a) - f_1^t(h^{t-1})(a)| > \varepsilon$  implies that condition (1) holds. Consequently, the process  $\{\tilde{L}_k\}$  omits at most those observations where  $|f_0^t(h^{t-1})(a) - f_1^t(h^{t-1})(a)| \leq \varepsilon$  for all  $a$  in  $A$ .

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

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

The key point is that  $K$  depends only on  $\varepsilon$  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  $1 - \varepsilon$ , either  $|f_0^t(h^{t-1})(a) - f_1^t(h^{t-1})(a)| < \varepsilon$  for all  $a$  in all but at most  $K$  periods or  $L_n < 1$ .

<sup>14</sup>See the Appendix where their result on the rate of convergence of supermartingales is formally stated. Our use of their result is similar to their main reputation theorem, but is sufficiently different that it is necessary to replicate parts of their argument. In the reputation context, the key object is the short-run player's forecast of the behavior of the long-run player. The likelihood ratio and the implied belief about types are intermediate steps. In our theorem, on the other hand, the likelihood ratio is the primary object.

If Expert 1 uses a mixed strategy  $\varphi_1$ , the conclusion follows from Fubini's theorem applied to the product measure  $P \times \varphi_1$ , because (1)  $T^n$  is jointly measurable and (2)  $K$  is uniform over all forecasting strategies. *Q.E.D.*

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.<sup>15</sup> 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  $H$ , that is, 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  $\varepsilon$ , 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  $\varepsilon$ . The inconclusive verdict indicates an insufficient difference between the two experts for a statistically significant comparison at level  $\varepsilon$ .

#### 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.

##### 4.1. A Bayesian Game

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

- Nature chooses an element  $P \in \Delta(A^n)$  according to a probability distribution  $\mu$ .
- Expert 0 is informed of  $P$ , while Expert 1 only knows  $\mu$ .
- 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

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

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

- The payoff of Expert 0 is  $1 - T^n(a^n, f_0, f_1)$ .
- Payoffs are extended to mixed strategies in the usual way.

<sup>15</sup>We thank Yossi Feinberg for pointing out this possibility.

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  $\mu$  is. For example, if  $\mu$  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  $\mu$  is even slightly diffuse, then his value is low when the horizon is long enough.

First, we define our notion of diffuseness. Any randomization by Nature  $\mu$ , being a distribution on  $\Delta(A^n)$ , extends to a distribution  $\bar{\mu}$  on  $\Delta(A^n) \times A^n$  via the formula

$$\bar{\mu}(Z) \equiv \int_{\Delta(A^n)} P(\{a^n : (P, a^n) \in Z\}) d\mu(P)$$

for every measurable  $Z \subset \Delta(A^n) \times A^n$ . Define  $\mathcal{M}(\varepsilon, \delta, L) \subset \Delta(\Delta(A^n))$  to consist of all  $\mu$  such that there are at least  $L$  periods  $t, 1 \leq t \leq n$ , such that

$$(2) \quad \max_{p \in \Delta(A)} \bar{\mu}^t(B_\varepsilon(p) | a^{t-1}, \alpha^{t-1}) < 1 - \delta, \quad \bar{\mu}\text{-a.e. } h^n,^{16}$$

where  $\bar{\mu}^t$  denotes the one-step-ahead conditional under  $\bar{\mu}$ . Note that the condition defining  $\mathcal{M}(\varepsilon, \delta, L)$ , which states that in each of at least  $L$  periods  $\mu$  does not concentrate its mass in some small ball, becomes less restrictive as  $n$  becomes large.<sup>17</sup>

In the **proof** of Theorem 2 the informed expert is assumed to report the truth, since his value without this constraint can only be higher. This motivates our definition of  $\mathcal{M}$ , in that the conditional expectation in (2) is the belief of the uninformed expert at time  $t$ , assuming that the informed expert reports the truth.

**THEOREM 2:** *For every  $\varepsilon$  and  $\delta > 0$  there is an integer  $L$  such that for every  $\mu \in \mathcal{M}(\varepsilon, \delta, L)$  the value of the game to Expert 1 is less than  $\varepsilon$ .*<sup>18</sup>

**PROOF:** Assume that the informed expert reports the truth. Let  $K = K(\varepsilon/2)$  be the integer obtained in Theorem 1. Let  $L = L(\varepsilon, \delta)$  be the smallest integer

<sup>16</sup>The notation  $B_\varepsilon(p)$  denotes the  $\varepsilon$  ball around  $p$ , where  $\Delta(A)$  is given the “max” norm it inherits as a subset of  $\mathcal{R}^{\#A}$ .

<sup>17</sup>For  $n < L$ , the set  $\mathcal{M}(\varepsilon, \delta, L)$  is empty.

<sup>18</sup>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.

so that the binomial distribution with  $L$  trials and probability  $\delta$  assigns probability at most  $\frac{\varepsilon}{2}$  to  $\{0, \dots, K\}$ .

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

Theorem 1 tells us that when the above case does not obtain, Expert 1’s probability of winning is at most  $\frac{\varepsilon}{2}$ . We conclude that his overall winning probability is at most  $\varepsilon$ . *Q.E.D.*

### 4.3. The Nonmanipulability 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  $\varepsilon$ .

**COROLLARY 3:** *For every  $\varepsilon$  and  $\delta > 0$  there is an integer  $L$  such that for every  $\mu \in \mathcal{M}(\varepsilon, \delta, L)$  and every  $\varphi_1$  there is  $P \in \text{supp } \mu$  such that*

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

**PROOF:** From Theorem 2 we have, for any such  $\mu$ ,

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

Then there must be an element in  $P \in \text{supp } \mu$  such that the conclusion of the theorem holds. *Q.E.D.*

### 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:

- $\mu_1$  is characterized by  $\bar{\mu}_1^t(\cdot|\alpha^{t-1})$  being the uniform distribution, independently across partial histories, on the vertices of  $\Delta(A)$ .
- $\mu_2$  is characterized by  $\bar{\mu}_2^t(\cdot|\alpha^{t-1})$  being the uniform distribution, independently across partial histories, over a small ball around the distribution  $\bar{p}$  that assigns equal probability to all outcomes.
- $\mu_3$  is defined similarly, except that  $\bar{\mu}_3^t(\cdot|\alpha^{t-1})$  puts unit mass on  $\bar{p}$ .

Fix a sufficiently large  $n$  so that  $\mu_1$  and  $\mu_2$  defined above both belong to  $\mathcal{M}(\varepsilon, \delta, L)$  for some  $\varepsilon, \delta > 0$  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

$\mu_1$  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  $\mu_2$  or  $\mu_3$  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  $\mu_3$ , both players are uninformed, and so they achieve equal value of 0.5. Under  $\mu_2$ , 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 varies 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 probabilities. What if Expert 0 has only partial, rather than exact, knowledge of the true distribution? We note here that this case can be adapted into our framework.

Modify the model of Section 4.1 by assuming that Expert 0's knowledge is given by a finite partition  $\Pi$  on  $\Delta(A^n)$  together with the prior  $\mu \in \Delta(\Delta(A^n))$ . Assume that  $\mu(\pi) > 0$  for each  $\pi \in \Pi$  and observe that Expert 0's belief about  $A^n$  upon observing  $\pi \in \Pi$  is given by  $P_\pi = \frac{1}{\mu(\pi)} \int_\pi P d\mu$ .

This is a model with *partial information*, where Expert 0 does not know the true probability  $P$ , but only the partition element  $\pi$  to which  $P$  belongs.<sup>19</sup> However, this model is equivalent to a *quotient model*, where the set of possible distributions is  $\{P_\pi : \pi \in \Pi\}$  with prior given by  $\mu'(P_\pi) = \mu(\pi)$ . Expert 0 in the quotient model knows the true distribution  $P_\pi$ , yet players' strategy sets and payoffs are equivalent to those in the partial information model. Both Theorems 1 and 2 apply to the quotient model.

## 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. Our model readily extends to the infinite-horizon case, and most of our results also extend—in fact in a stronger form.<sup>20</sup>

<sup>19</sup>We continue to assume that Expert 1 has no information about  $P$  (beyond  $\mu$ ).

<sup>20</sup>The details are standard: In the infinite horizon, the sets of infinite realizations  $A^\infty$  and histories  $H^\infty$  are given the product topologies. Probabilities on these spaces are defined on the

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  $\limsup_{n \rightarrow \infty} L_n(h^n) < 1$ , Expert 1 if  $\liminf_{n \rightarrow \infty} L_n(h^n) > 1$ , and 0.5 otherwise. The constant  $K$  derived in Theorem 1 is independent of the horizon.<sup>21</sup>

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  $\varphi_j$  ( $j \neq i$ ), there is  $P \times \varphi_j$  probability 1 that either*

- (a)  *$T$  picks expert  $i$  or*
- (b)  *$\lim_{t \rightarrow \infty} |f_0^t(h^{t-1}) - f_1^t(h^{t-1})| = 0$ .*

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

$$A_n \equiv \left\{ h^\infty : \limsup_{t \rightarrow \infty} L_t \geq 1 \ \& \ \#\{t : |f_0^t(h^{t-1}) - f_1^t(h^{t-1})| > \varepsilon_n\} > K_n \right\}$$

has probability less than  $\varepsilon_n$ .<sup>22</sup> Since  $\sum_n P(A_n) < \infty$ , by the Borel–Cantelli lemma we have

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

Thus, for  $P$ -a.e. path  $h^\infty$ , either Expert 0 wins or, for all but finitely many  $n$ ,  $|f_0^t(h^{t-1}) - f_1^t(h^{t-1})| \leq \varepsilon_n$  for all but finitely many  $t$ . In the latter case  $|f_0^t(h^{t-1}) - f_1^t(h^{t-1})| \rightarrow 0$ . *Q.E.D.*

Our final result shows that Theorem 2 extends to the infinite horizon in a sharper form. Any  $\mu \in \Delta(\Delta(A^\infty))$  defines a  $\bar{\mu}$  as in Section 4.1. Let  $\mathcal{M}(\varepsilon, \delta) \subset \Delta(\Delta(A^\infty))$  be the set consisting of all  $\mu$ 's such that for  $\mu$ -a.e. infinite history  $h^\infty$ , for infinitely many periods,

$$(3) \quad \max_{p \in \Delta(A)} \bar{\mu}^t(B_\varepsilon(p) | \alpha^{t-1}) < 1 - \delta.$$

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

Borel  $\sigma$ -algebras on these spaces. Mixed strategies are defined on the Borel  $\sigma$ -algebra generated by the weak topology on  $\Delta(A^\infty)$ .

<sup>21</sup>As it is in the FL active supermartingale result.

<sup>22</sup>The argument in the proof of Theorem 1 can be readily cast in an infinite-horizon setting to draw the conclusion that for every  $\varepsilon$  there is an integer  $K$  such that with  $P$  probability at least  $1 - \varepsilon$ , either (a)  $\limsup L_n < 1$  or (b) the two-experts' forecasts are  $\varepsilon$ -close in all but  $K$  periods.

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

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

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

$$\bar{\mu}\{(f_0, h^\infty) : |f_0^t(h^{t-1}) - f_1^t(h^{t-1})| > \varepsilon \text{ i.o.}\} = 1,$$

so case (b) in Theorem 4 holds with probability 0. The payoff of the strategic expert is therefore 0. *Q.E.D.*

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 emphasize the infinite horizon, while our focus is on putting bounds on the errors in a finite horizon. They also have a different framework for evaluating 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 specification of a distribution over distributions to represent the false expert's uncertainty about the true probabilities. By contrast, a classical decision-maker evaluates the subjective probability of passing rather than the category of the set on which he passes. This is the motivation for our introduction of second-order distributions in Section 4.

## 6. DISCUSSION

We begin with an informal review of [Sandroni's \(2003\)](#) disarmingly elegant use of the minimax theorem to prove impossibility.

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. Assume that  $n$  is finite.

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

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

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:

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

Extend  $z_s$  to mixed strategies  $\mu$  and  $\varphi$  in the usual way.

The impossibility result asserts that the expert has a strategy  $\varphi$  that guarantees him a high payoff regardless of what Nature does. To prove the result, think of the forecaster as playing a constant-sum game against Nature, so that the minimax theorem asserts

$$(4) \quad \max_{\varphi \in \Delta(\Delta(A^n))} \min_{\mu \in \Delta(\Delta(A^n))} z_s(\mu, \varphi) = \min_{\mu \in \Delta(\Delta(A^n))} \max_{\varphi \in \Delta(\Delta(A^n))} z_s(\mu, \varphi).$$

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

The crucial observation is that Nature's randomization is completely superfluous. Let  $P^\mu$  denote the probability measure obtained from  $\mu$  by the reduction of compound lotteries. As far as the payoffs are concerned, whether Nature uses a mixed strategy  $\mu$  or  $P^\mu$  makes no difference:

$$(5) \quad z_s(\mu, \varphi) = z_s(P^\mu, \varphi) \quad \forall \mu, \varphi \in \Delta(\Delta(A^n)).$$

This is because  $\mu$  and  $P^\mu$  induce identical distributions on the set of outcomes  $A^n$ . As far as realized outcomes are concerned,  $\mu$  and  $P^\mu$  are observationally indistinguishable. For example, outside observers 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  $P^\mu = (P^1 + P^2)/2$ . By contrast, an expert's mixed strategy  $\nu$  is not, in general, reducible: choosing between the two forecasts  $Q^1$  or  $Q^2$  with equal probability is not payoff equivalent to the forecast  $Q = (Q^1 + Q^2)/2$ .

Given this asymmetry between Nature's and the expert's randomizations, the conclusion of the minimax theorem can be rewritten as

$$(6) \quad \max_{\varphi \in \Delta(\Delta(A^n))} \min_{P \in \Delta(A^n)} z_s(P, \varphi) = \min_{P \in \Delta(A^n)} \max_{\varphi \in \Delta(\Delta(A^n))} z_s(P, \varphi).$$

Here is where the assumption that a test  $T_s^n$  passes the truth with probability  $1 - \varepsilon$  plays its critical role. This assumption, which states that for all  $P$ ,

$$(7) \quad z_s(P, P) \equiv P\{T_s^n(a^n, P) = 1\} > 1 - \varepsilon,$$

ensures that the right-hand side of Eq. (6) is at least  $1 - \varepsilon$ . If the expert knows that Nature has chosen  $P$ , then he has an obvious response guaranteeing a payoff of  $1 - \varepsilon$ , namely to report  $P$ . This delivers the conclusion that the maximin value is also greater than  $1 - \varepsilon$ , that is, the strategic expert can pass the test with high probability.

To sum up, the key to understanding the impossibility theorem is the reducibility of Nature's compound lotteries in the sense of Eq. (5) above. This reducibility means that the seemingly innocuous assumption that the expert has a good response to any pure strategy  $P \in \Delta(A^n)$  also implies he has a good response to any mixed strategy  $\mu \in \Delta(\Delta(A^n))$ . This allows the power of the minimax theorem to come into play, delivering the desired result. Thus, the expert can win a hide-and-seek game where Nature hides the true probability  $P$ , despite the large number of potential hiding places, because in the single-expert setting Nature has no meaningful opportunity to randomize.

Our results on comparative testing may be understood as a consequence of the restoration of  $\Delta(\Delta(A^n))$  as Nature's strategy space. Consider again the game in Section 4, where Nature uses a mixed strategy  $\mu$  and informs Expert 0 of its random choice  $P \in \Delta(A^n)$ . The presence of an informed expert breaks the strategic equivalence between  $\mu$  and  $P^\mu$ . Unless  $\mu$  is degenerate, Nature's use of a mixed strategy  $\mu$  is now strategically distinct from  $P^\mu$ , in the sense that Eq. (5) no longer holds. To win, the strategic expert must, at least approximately (in the sense of Theorem 1), guess Nature's selection of a pure strategy  $P$ . This he cannot guarantee. 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.

Where does that leave us with the assumption that a test must pass the truth and the notion of stochastic truth itself? 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  $\mu$  (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, such as Sandroni's (2003) theorem, provide invaluable insights by uncovering the subtle consequences of their assumptions. 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 drives the impossibility result 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.<sup>23</sup> There are 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 many 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 there is no superior alternative.<sup>24</sup> 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 their peers and 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.

*Dept. of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston, IL 60208, U.S.A.; al-najjar@northwestern.edu; <http://www.kellogg.northwestern.edu/faculty/alnajjar/htm/index.htm>*

*and*

<sup>23</sup>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 expert can produce a falsifiable theory  $Q$  that is unlikely to be rejected by that rule, regardless of what the truth is. Harman and Kulkarni (2007) provided a different perspective and discussed the limitations of simplistic Popperian falsifiability when theories are probabilistic.

<sup>24</sup>For details on these examples, see Darling (2006).

*Dept. of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston, IL 60208, U.S.A.; j-weinstein@kellogg.northwestern.edu; <http://www.kellogg.northwestern.edu/faculty/weinstein/htm/index.htm>.*

*Manuscript received January, 2007; final revision received January, 2008.*

## APPENDIX

### A.1. The Active Supermartingale Theorem

Consider an abstract setting with a probability measure  $P$  on  $H^\infty$  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) showed the following remarkable result:

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

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

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

The power of the theorem stems from the fact that the integer  $K$ , which depends on the parameters  $l_0$ ,  $\varepsilon$ ,  $\psi$ , 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) introduced 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

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

with the interpretation that  $\tau(a^n, f_0, f_1) = 1$  if and only if at least one expert is informed.<sup>25</sup>

The following theorem is an important variant of Sandroni’s (2003) impossibility result:

THEOREM 7: *Suppose that  $\tau$  is such that for every  $P, f_0$ , and  $f_1$*

$$(8) \quad P\{a^n : \tau(a^n, f_0, f_1) = 1\} > 1 - \varepsilon \quad \text{if either } f_0 = f_P \text{ or } f_1 = f_P.$$

*Then, for every mixed strategy  $\varphi_0$  of Expert 0, there is a mixed strategy  $\varphi_1$  of Expert 1 such that, for every  $a^n$ ,*

$$(9) \quad \varphi_0 \times \varphi_1\{(f_0, f_1) : \tau(a^n, f_0, f_1) = 1\} > 1 - \varepsilon.$$

That is, if  $\tau$  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

<sup>25</sup>Note that we allow the test  $\tau$  to condition on the entire forecasting schemes, including forecasts at unobserved histories. This only strengthens the conclusion of Theorem 7.

opponent strategy, manipulate  $\tau$  by forcing it to return 1 (with high probability) without any knowledge of the true process.

PROOF OF THEOREM 7: For any forecasting strategy  $f_0$  of Expert 0, define the single-expert test

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

by

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

By (8), the single-expert test  $M_{f_0}$  passes the truth with probability  $1 - \varepsilon$ . From Sandroni (2003) we know that there is a mixed strategy  $\varphi_1$  such that for every  $a^n$ ,

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

This establishes (9) for pure  $\varphi_0$ .

For a general  $\varphi_0$ , Expert 1 is facing a lottery over deterministic tests. We show that Sandroni’s (2003) 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  $M_{\varphi_0}$  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  $f_1$ ,

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

Applying the Minimax Theorem (Fan (1953)), we conclude that there is  $\varphi_1$  such that, for every  $a^n$ ,

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

from which (9) directly follows.

*Q.E.D.*

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 \(2007\)](#) proved an impossibility theorem for all infinite-horizon tests that do not use counterfactuals.

## REFERENCES

- DARLING, D. (2006): *Gravity's Arc*. New York: Wiley. [555]
- DEKEL, E., AND Y. FEINBERG (2006): "Non-Bayesian Testing of an Expert," *Review of Economic Studies*, 73, 893–906. [542,552]
- FAN, K. (1953): "Minimax Theorems," *Proceedings of the National Academy of Sciences of the United States of America*, 39, 42–47. [558]
- FEINBERG, Y., AND C. STEWART (2006): "Testing Multiple Experts," *Econometrica*, 76, 561–582. [542,552]
- FOSTER, D., AND R. VOHRA (1998): "Asymptotic Calibration," *Biometrika*, 85, 379–390. [541]
- FUDENBERG, D., AND D. K. LEVINE (1992): "Maintaining a Reputation When Strategies Are Imperfectly Observed," *Review of Economic Studies*, 59, 561–579. [542,546,556]
- (1999): "An Easier Way to Calibrate," *Games and Economic Behavior*, 29, 131–137. [542]
- HARMAN, G., AND S. KULKARNI (2007): *Reliable Reasoning: Induction and Statistical Learning Theory*. Cambridge, MA: MIT Press. [555]
- KALAI, E., AND E. LEHRER (1994): "Weak and Strong Merging of Opinions," *Journal of Mathematical Economics*, 23, 73–86. [556]
- KALAI, E., E. LEHRER, AND R. SMORODINSKY (1999): "Calibrated Forecasting and Merging," *Games and Economic Behavior*, 29, 151–159. [542]
- LEHRER, E. (2001): "Any Inspection Is Manipulable," *Econometrica*, 69, 1333–1347. [542]
- OAKES, D. (1985): "Self-Calibrating Priors Do Not Exist," *Journal of the American Statistical Association*, 80, 339–339. [548]
- OLSZEWSKI, W., AND A. SANDRONI (2006): "Strategic Manipulation of Empirical Tests," Report, Northwestern University. [542,543,552]
- (2007): "Future-Independent Tests," Report, Northwestern University. [542,559]
- SANDRONI, A. (2003): "The Reproducible Properties of Correct Forecasts," *International Journal of Game Theory*, 32, 151–159. [541,542,552,554,557,558]
- SANDRONI, A., R. SMORODINSKY, AND R. VOHRA (2003): "Calibration With Many Checking Rules," *Mathematics of Operations Research*, 28, 141–153. [542]
- SORIN, S. (1999): "Merging, Reputation, and Repeated Games With Incomplete Information," *Games and Economic Behavior*, 29, 274–308. [557]