Comparative Testing of Experts*

Nabil I. Al-Najjar†
and
Jonathan Weinstein‡

First draft: November 2006
This version: January 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 an elegant result due to Fudenberg and Levine (1992) on the convergence of reputations.

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, Wojciech Olszewski, 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 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 either

1. the test will pick the informed expert; or

2. the uninformed expert makes forecasts that are close to the truth in most periods.

[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.

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 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 3 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 should be emphasized that our comparative test does not blunt the force of the single-expert impossibility results. In Theorem 2 we show that there can be no non-manipulable test that can tell when there is at least one informed expert. This shows that any positive results in multiple-expert settings must be of the form: if there is an informed expert, the test will select him.

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 based on a single sample path. 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 $i = 1, 2$ 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 $\big[A, \Delta(A), \Delta(A)\big]^t$.

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 strategies that coincides with the one-period-ahead conditionals at partial histories 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]

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

[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.

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]

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

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

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

Here, $i = T^n(h^n)$ is interpreted to mean that the test picks expert i after observing the history of forecasts and Nature's realizations for the past n periods.

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 of any kind: What the experts would have forecasted at unrealized histories is not taken into account; only forecasts along the actual history are used;

3 A Comparative Test of Experts

An expert is *informed* if he forecasts outcomes using the true distribution P. Formally, his strategy is the deterministic forecast f_P. In Theorem 2 we will show that no test can determine whether or not at least one of the two experts is informed. Therefore the appropriate goal is a comparative test that picks an informed expert if there is indeed one.

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

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

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

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 an expert is chosen at random with probability 0.5 if $L_n(h^n) = 1$.

Theorem 1 *For every $\epsilon > 0$, there is an integer K such that for all integers n, distributions P, and mixed forecasting strategies φ_0, φ_1 with at least one informed expert, there is P-probability at least $1 - \epsilon$ that either*

(a) *T^n picks an informed expert; 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 correctly guess the true law P. Note that the theorem has no bite when n is small relative to 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 3 will support the conclusion that case (b) is "unlikely" when n is large relative to K.

Proof: The argument relies on a result by Fudenberg and Levine (1992) on the rate of convergence of supermartingales.[9]

Assume for the moment that Expert 0 is informed and that he reports the truth.[10] It is a standard observation that the stochastic process $\{L_t\}$ is a

[9] Although our test has a "reputation" flavor and we use some of Fudenberg and Levine (1992)'s techniques, the analogy between the two frameworks is not transparent. For the reader familiar with their paper, one may think of the realized outcomes in our model as the random signals generated by the long-run player's actions. The two experts' forecasts are then the distributions over signals induced by the mixed strategies of the two types of this player. The tester, who updates beliefs over two experts, is analogous to the short-run players, who update their beliefs over two types.

[10] The informed expert may have a strategy that does better than reporting the truth; if so, this only strengthens our conclusion.

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 $\{\tilde{L}_t\}$ to be the faster process obtained from $\{L_t\}$ through a sequence of stopping times that contains all finite histories at which $|f_0^t(h^{t-1})(a(t)) - f_1^t(h^{t-1})(a(t))| > \epsilon$.

Fudenberg and Levine show that $\{\tilde{L}_t\}$ is an active supermartingale with activity ϵ. We refer the reader to their paper for definitions. Their Theorem A.1 implies that for any $\epsilon > 0$ there is an integer K such that for any active supermartingale $\{\tilde{L}_t\}$

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

The key point is that K depends only on ϵ 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 informed, with probability $1 - \epsilon$, on any history of n periods, either $|f_0^t(h^{t-1})(a(t)) - f_1^t(h^{t-1})(a(t))| < \epsilon$ for 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 assumption that T^n is jointly measurable and the fact that the constant K is uniform over all forecasting strategies. ■

To further elucidate the second part of the conclusion of the theorem, suppose that $A = \{Heads, Tails\}$ and P is an i.i.d. distribution with probability of Heads α. Assume that the strategic expert knows that P is i.i.d., but does not know the value α. If this expert estimates the true value of α from the data then whether or not he will be picked will depend on how fast he comes close to learning α relative to the size of K. Unfortunately, useful bounds on the value of K are not known, but if the true value of K happens to be large, then the expert who only knows the process is i.i.d. may end up being picked. Note, however, that such an expert is hardly uninformed; after all, he knows that the true distribution belongs to a simple one-parameter family and he eventually forecasts outcomes almost as well as the informed expert.

Now we turn to 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. The following theorem is an important consequence of Sandroni (2003)'s impossibility result:

Theorem 2 *Suppose that τ is such that for every P, f_0 and f_1*

$$P\{a^n : \tau(a^n, f_0, f_1) = 1\} > 1 - \epsilon \quad \text{if either } f_0 = f_P \text{ or } f_1 = f_P. \tag{2}$$

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

$$\varphi_0 \times \varphi_1\big\{(f_0, f_1) : \tau(a^n, f_0, f_1) = 1\big\} > 1 - \epsilon. \tag{3}$$

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

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 2, 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

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

This establishes 3 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

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

from which 3 directly follows. ■

4 The Scope of Strategic Manipulations

Theorem 1 establishes statistical properties of a simple "reputation-style" test, taking the experts's 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 $n = 1, 2\ldots$ 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:

$$z(\mu, \varphi) \equiv \int_{\Delta(A^n)} \int_{F^n} \left[\int_{A^n} T^n\big(a^n, f_0, f_1\big)\, dP(a^n) \right] d\varphi(f_1)\, d\mu(P). \quad (4)$$

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

$$\max_{p \in \Delta(A)} \mu^t(B_\epsilon(p)|\alpha^{t-1}) < 1 - \delta.^{11} \quad (5)$$

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

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 3 *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.*

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 6, 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. ■

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 4 *For every φ_1 and every $\mu \in \mathcal{M}(\delta_1, \delta_2, L)$ there is $P \in \operatorname{supp} \mu$ such that*

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

Proof: From Theorem 3 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 3.

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.

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 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.[12] The constant K derived in Theorem 1 is independent of the horizon.[13]

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 5 *For any distribution P and mixed forecasting strategies φ_0, φ_1 with at least one informed expert, with P-probability 1 either*

(a) *T picks an informed expert; 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 informed, 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

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

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

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

has probability less than ϵ_n.[14, 15] 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 $|f_0^t(h^{t-1}) - f_1^t(h^{t-1})| \to 0$. ■

Theorem 2, by contrast, does *not* extend to the infinite horizon 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 which, like ours, do not use counterfactuals. This again provides a contrast between the single- and multiple-expert frameworks.

Our final result shows that Theorem 3 extends to the infinite horizon in a sharper form. The Bayesian game in Section 4.1 extends to the infinite horizon setting.[16] 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,

$$\max_{p \in \Delta(A)} \mu^t(B_\epsilon(p)|\alpha^{t-1}) < 1 - \delta. \tag{6}$$

Theorem 6 *For every ϵ, $\delta > 0$ and $\mu \in \mathcal{M}(\epsilon, \delta)$ the value of the game to Expert 1 is zero.*

[14] We note that it is evident from the proof of Theorem 1 that finiteness of the horizon is 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.

[15] By appealing to Footnote 12, we ignore the possibility that L_t might not converge.

[16] 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 $\Delta(A^\infty)$.

Proof: The proof closely follows that of Theorem 3 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

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

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 5 holds with probability 0. The payoff of the strategic expert is therefore 0. ■

We now discuss the recent interesting and independent work of Feinberg and Stewart (2006). Feinberg and Stewart study an infinite horizon model of testing multiple experts using a cross-calibration test. Although the motivation is similar, the results are quite distinct. The first obvious difference is that our tests are motivated by very different considerations (reputation vs. calibration) and thus work quite differently. Second, the finite-horizon setting we work with has some advantages, including its relevance to applications where the test is used as basis for a decision to be made at a given point in time. Also, the finite-horizon setting is conceptually clearer since the impossibility result of Sandroni (2003) holds without any qualifications, unmarred by technical issues arising in the infinite horizon case.[17] Third, in Theorem 1 of their paper they show that two uninformed strategic experts cannot simultaneously pass the test (except on a category 1 set of probability measures). Our analysis says little about the case of two uninformed strategic experts. Finally, the conclusion of our Theorem 3 is in terms of the value in an incomplete-information constant-sum game, and has a simple probabilistic interpretation. Feinberg and Stewart (2006) show that in the

[17] These issues are especially delicate since discounting does not seem to be meaningful in the testing setting. Undiscounted infinite horizon problems are problematic because of the loss in the compactness of the strategy spaces.

presence of an informed expert their test is non-manipulable, in the sense that for every strategy of the uninformed expert there is a true distribution for which he fails with probability 1. They show in addition is that for every such strategy this expert will fail for "almost all" true distributions, where "almost all" is in the sense of all but category one set of distributions.

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:

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

The impossibility theorem boils down to putting a lower bound on 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:

$$z_s(P,P) \equiv P\{T_s^n(a^n,P)=1\} > 1-\epsilon. \tag{8}$$

This condition ensures that the RHS of Eq. 7 is close to 1: if the expert knew that Nature has chosen P, then he has an obvious best response, namely to report P, guaranteeing himself a payoff of $1-\epsilon$. This delivers the conclusion that it is impossible to design a test that a strategic expert cannot pass 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:

$$\max_{\varphi\in\Delta(\Delta(A^n))} \min_{\mu\in\Delta(\Delta(A^n))} z_s(\mu,\varphi) = \min_{\mu\in\Delta(\Delta(A^n))} \max_{\varphi\in\Delta(\Delta(A^n))} z_s(\mu,\varphi). \tag{9}$$

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:

$$z_s(\mu,\varphi) = z_s(P_\mu,\varphi), \quad \forall \mu,\varphi \in \Delta(\Delta(A^n)). \tag{10}$$

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 or 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. 7 and 9 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. 10 no longer holds. To win, the the strategic expert has to guess Nature's selection of a pure strategy. What changed relative to the single-expert model is that the presence of the informed expert's forecasts breaks the observational equivalence between μ and P_μ.

We should emphasize that the issue is not that Nature does not have the opportunity to randomize, but whether randomization is meaningful in terms of payoffs (Eq. 10). When randomization is superfluous, as in Eq. 7, the expert is "a step ahead," giving him the advantage against Nature.

These observations provide a systematic way to understand why some structural assumptions are critical for the impossibility results. For example, why are they inconsistent with repeated sampling, so commonly assumed in statistical inference? Consider the variant of the single-expert model where the only departure is that we now provide the test with repeated samples generated independently by the same unknown distribution μ. With many such samples, one can find a test such that a strategic expert cannot ignorantly pass. The reason here is that Nature's strategy space is no longer

reducible to $\Delta(A^n)$: a μ that picks either P_1 or P_2 with equal probability generates observations according to either P_1 or P_2, and this is observationally distinguishable from observations generated by P_μ.

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 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 an expert's 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.[18] 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 various empirical anomalies—not to mention its implicit reliance on "action at a distance" in the transmission of gravitational force. 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.[19] Economics is full of similar examples. Expected utility theory continue 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.

The classical, frequentist, view attaches probabilities only to events that are subject to repeated identical trials. In the context of testing experts,

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

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

such repetition is not possible, since probability laws may change arbitrarily every period. The impossibility results can be seen as a confirmation of the classical view. If there is no effective test for the truth, perhaps the concept of "true" probabilities is not worthwhile.

To what extent does our comparative test recover the concept of truth? Using an extension of our test to compare multiple experts, we can say that if anyone knows the truth, our chosen expert does. If this is the best we can do, perhaps the appropriate interpretation is that for all practical purposes, the truth is always relative. We cannot say whether or not a theory is correct in an absolute sense, only that it is better than the others.

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 show that a very simple reputation-type comparative test provides both normatively and descriptively appealing method of testing experts.

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

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

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

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.

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.