Comparative Testing of Experts

Nabil I. Al-Najjar*
and
Jonathan Weinstein†

December 2006

Preliminary Draft

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 test relies on a simple reputation argument due to Fudenberg and Levine (1992).

We then 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 is a consequence of the fact that, in the single-expert setting, any mixed strategy for Nature is reducible to a pure strategy, thus eliminating any meaningful sense in which Nature can randomize. Comparative testing reverses the impossibility result because the presence of an informed expert eliminates the reducibility of Nature's mixed strategies.

* 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

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 the weather, stock returns, or GDP. 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), who showed that a particular class of test, known as calibration tests, can be passed by a strategic but totally ignorant expert.[1] Their main insight is that a strategic expert who knows nothing about the underlying process can pass a calibration test on *any* sample path. 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 the result to passing many calibration rules simultaneously, and 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 path of play. The remarkable feature of this result is that it is not limited to any special class of tests, and it requires only that the test is not so stringent that an expert who knows the truth cannot pass it.

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. Olszewski and Sandroni (2006b) refine these findings and obtain additional results. These tests, however, will not validate a true expert in finite time. Olszewski and Sandroni (2006a)

[1] 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.

show that the impossibility result is restored if one also requires a test not to condition on counterfactual predictions, namely forecasts at unrealized future histories.

In this paper we show that these impossibility results do not extend to tests that compare two (or more) experts. For expository clarity, we shall ignore quantifiers on probabilities and degrees of approximation in the introduction. We also confine our discussion to the finite horizon case where the issues are conceptually clearer, unmarred by technical considerations arising in infinite horizon settings. We believe, however, that the points we make extend naturally 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: 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.

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. Our second theorem shows that this is indeed the case: We compute an upper bound on the uninformed expert's value in a zero-sum game where Nature chooses a true P and informs the other expert. We use the first theorem to show that the uninformed expert's value can be made as small as one wishes if the (finite) horizon is long enough. Finally, our third result shows that when using our test with two partially informed experts, the expert with better information always does better.

The comparative testing setting explored in this paper makes a slightly more general point by shedding light on the source of the impossibility results. Roughly speaking, we argue that the impossibility results are con-

sequences 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 observation of the process. In the zero-sum game between an expert and Nature, 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 explanation why impossibility results require tests and auxiliary assumptions with seemingly odd properties. For instance, impossibility results are incompatible with tests that reward information (in a sense we make precise), with repeated observations of the stochastic process, or with comparison across experts, as we do here. All of these variants either fully or partially restore the richness of Nature's strategy set and they are consequently incompatible with the impossibility result. Section 5 elaborates on these points extensively.

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 $\alpha_i(t)$, $i = 0, 1$, is the probabilistic forecast of expert i for that period.[2] 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. It is standard to show that any such strategy defines a unique stochastic process, where we interpret f^t as the one-step-ahead conditionals. Let F^n denote the set of all forecasting strategies.

We shall think of F^n as an expert's set of pure strategies. Experts may

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

randomize by choosing $\varphi \in \Delta(F^n)$ representing a probability distribution φ on pure strategies. [3]

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).

There is a true stochastic process P on A^n that generates outcomes. Note that any stochastic process defines a forecasting strategy as one-period-ahead conditionals at partial histories with positive probability, and setting forecasts on zero-probability events arbitrarily.

An *n-period comparative test* is any measurable function[4]

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

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;
- The test does not condition on counterfactuals of any kind: What the experts would have forecasted at unrealized histories is not taken into account; the test only uses the experts' forecasts along the actual history;
- 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 that expert's label.

[3] 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.

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

3 A comparative test of experts

An expert is *informed* if he forecasts outcomes using the true distribution. Formally, his forecasting strategy φ puts unit mass on the deterministic forecast f in which, for every action a, $f^t(h^{t-1})(a) = P\big(a(t) = a|a^{t-1}\big)$ for any history h^{t-1} whose outcome component a^{t-1} has positive probability under P.

We say that 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.

Theorem 1 *Fix $\epsilon > 0$. There is an integer K such that for all n there is an n-period comparative test T^n such that for any distribution P and any mixed forecasting strategies φ_0, φ_1 with at least one informed expert then, with P-probability $1 - \epsilon$, 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 of what the true law is. We will subsequently confirm (Theorem 2) that case (b) is "unlikely" to obtain when n is large relative to K.

Proof: The idea is to convert the game to a Bayesian game in which the test compares the "reputation" of the two experts, choosing the one with the highest posterior odds of being a better forecaster. We introduce a stochastic process L_t reflecting the evolution of the odds ratio (the ratio of the experts' scores, or the probability that each is informed) as a function of the realized outcome and the experts' forecasts. The test begins by assigning an initial score 0.5 to each expert, which we shall interpret as the prior probability that the expert is a better forecaster.

Define $L_0(h^0) = 1$ and let

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

be the updating rule for the odds ratio along a history h^n.

The argument relies on a result by Fudenberg and Levine (1992) (henceforth denoted FT) on the rate of convergence of supermartingales. Assume for the moment that Expert 0 is informed and that he reports the truth.[5] It is a standard observation that the stochastic process $\{L_t\}$ is a supermartingale under the distribution induced by the strategy of Expert 0 (FL, Lemma 4.1). As in FL, 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 $|f_0^t(h^{t-1})(a(t)) - f_1^t(h^{t-1})(a(t))| > \epsilon$.

FL 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 $\epsilon > 0, z \in (0,1)$ 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 z, and not on the true stochastic process P or the forecasting strategy f_1.

We can now define the test: 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 if $L_n(h^n) = 1$.

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 $|f_0^t(h^{t-1})(a(t)) - f_1^t(h^{t-1})(a(t))| < \epsilon,\ t \leq K$ 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.

Finally, we note that if there are two informed experts, then the conclusion of the theorem is trivial. ■

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

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.

What happens if neither experts is informed? We suspect, but we haven't verified, that the theorem is "continuous" in the sense that if its setting is slightly perturbed by assuming that the informed expert's knowledge is not exact that its conclusion continues to hold. A more interesting, and more difficult enterprise, is to show that there are simple test like ours that always pick the most informed expert. Part of the difficulty may be finding the right definition for "most informed."

Notwithstanding this issue, one should not lose sight of the bottom line of the impossibility results: A test that passes an informed expert can be passed by an ignorant expert. Here we point out that this is an issue only when an informed expert and an ignorant expert cannot be directly compared.

4 A Game Against Nature

As indicated earlier, one cannot rule out in the conclusion of Theorem 1 the possibility that a completely uninformed expert might make a lucky guess that lands him close to P. Intuitively, however, the odds of this happening ought to be remote.

To make this formal, consider the following sequence of zero-sum games, where the nth game, $\Gamma(n)$, has the following properties:

- Two players, Nature and Expert 1, move simultaneously;

- T^n is the odds ratio test constructed in Theorem 1;
- Nature's set of pure strategies is $\Delta(A^n)$, and Expert 1's is F^n;
- Expert 0 reports the correct forecast f_{0P} derived from Nature's choice P;
- The payoff of Expert 1 is equal to the probability of being chosen:

$$z(P, f_1) \equiv \int_{H^n} T^n\big(h^n, f_{0P}, f_1\big)\, dP(h^n). \tag{2}$$

  Nature's payoff is $-z$.

This is a game between Nature and Expert 1 since Expert 0 is completely non-strategic. Both Nature and Expert 1 may play mixed strategies. An expert's mixed strategy is generically a $\varphi \in \Delta(F^n)$, while a mixed strategy by Nature is generically a $\mu \in \Delta(\Delta(A^n))$. Expected payoffs are defined in the obvious way, namely:

$$z(\mu, \varphi) \equiv \int_{\Delta(A^n)} \int_{F^n} \left[\int_{H^n} T^n\big(h^n, f_0, f_1\big)\, dP(h^n) \right] d\varphi(f_1)\, d\mu(P). \tag{3}$$

Informally, the next theorem 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, there is a distribution $P \in \Delta(A^n)$ for Nature such that the uninformed expert passes the test with probability at most ϵ.

Theorem 2 *For every $\epsilon > 0$ there is an integer n such that the value of the zero-sum game $\Gamma(n)$ to Expert 1 is no more than ϵ.*

The reader should recall that the test is symmetric, so there is no presumption that the tester knows that it is Expert 0 who is informed. Indeed the same theorem holds if the experts' roles are switched.

Proof: We apply the Minimax Theorem to this problem. A classic reference is Fan (1953). The strategy sets $\Delta(\Delta(A^n))$ and $\Delta(F^n)$ are compact (in the weak topology, which we assume throughout). From Eq. 3, it is clear that

z is continuous in both of its arguments. And since z is defined in terms of integrals, it is linear in μ given φ, and linear in φ given μ. This verifies the conditions of the Minimax Theorem, which implies:

$$\max_{\varphi\in\Delta(F^n)} \min_{\mu\in\Delta(\Delta(A^n))} z(\mu,\varphi) = \min_{\mu\in\Delta(\Delta(A^n))} \max_{\varphi\in\Delta(F^n)} z(\mu,\varphi).^{6} \tag{4}$$

Clearly,

$$\min_{\mu\in\Delta(\Delta(A^n))} \max_{\varphi\in\Delta(F^n)} z(\mu,\varphi) \leq \max_{f\in F^n} z(\bar{\mu},\varphi),$$

where $\bar{\mu}$ is the mixed strategy characterized by the following process: for each partial history h^{t-1}, a one-step-ahead conditional $p(h^{t-1})$ is chosen according to the uniform distribution on $\Delta(A)$, independently across partial histories. The resulting system of one-step-ahead conditionals defines uniquely an element of $\Delta(A^n)$, and thus the process of uniform independent draws defines an element $\bar{\mu} \in \Delta(\Delta(A^n))$.

On the other hand, to fall within an ϵ-neighborhood of $p(h^{t-1})$, the optimal pure strategies consist of the forecasts $f_1^t(h^{t-1}) \in \Delta_\epsilon(A)$, where the latter expression denotes the subsets of $\Delta(A)$ such that each of its entries is at least ϵ. Any such choice guarantees only a probability equal to the $\bar{\mu}$-measure of any ϵ ball contained within the interior of $\Delta(A)$. Denote this probability by α. Call a pure strategy f_1 optimal, if it selects an optimal forecast at each history.

Expert 1 can win only when case (b) of Theorem 1 holds. Since Nature is randomizing independently, for any n, any optimal pure strategy f_1 and along any history h^n, the probability that forecasts will be ϵ-close to the truth on at least $n-K$ periods is no greater than

$$\binom{n}{n-K} \alpha^{n-K}. \tag{5}$$

This is also true for any optimal mixed strategy, *i.e.,* one whose support consists of optimal pure strategies. The term $\binom{n}{n-K}$ is polynomial in n while α^{n-K} is exponential in n, so the expression in 5 converges to 0 as n goes to infinity (the fact that K does not depend on n is used to show that the polynomial is of fixed degree).

[6] Of the course, the inner min's and max's can be equivalently stated in terms of pure strategies. We do not do so for expository reasons.

Taking n large enough therefore guarantees that we can make

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

as small as we wish. ■

5 Discussion

For expositional clarity, we shall assume that the forecaster submits a measure Q, rather than just a forecast that is observed along the actual history. With this, a strategic expert's pure strategy set is $\Delta(A^n)$, and his mixed strategies are in $\Delta(\Delta(A^n))$, exactly the same as Nature's. Note that this, if anything, strengthens the tests as it reveals the expert's counterfactual predictions along unrealized histories.

5.1 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. His full argument is more involved; our description here provides just the intuition necessary to make our point.

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. The key tool

is the Minimax Theorem, whose assumptions can be readily verified, and asserts:

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

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:

$$z_s(P,P)\equiv P\{T_s^n(a^n,P)=1\}>1-\epsilon. \tag{7}$$

This condition ensures that the RHS of Eq. 6 is close to 1: if the expert knew that Nature has chosen P, then he has an obvious best response, namely to report a forecast P, which by the above requirement guarantees him 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 steps:

- The Minimax Theorem;
- A test must pass the truth.

We examine these two steps in turn.

5.2 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 proof of Theorem 2, we used an expression, Eq. 4, for the conclusion of the Minimax theorem that, in the single-expert setting, translates to:

$$\max_{\varphi\in\Delta(\Delta(A^n))}\ \min_{\mu\in\Delta(\Delta(A^n))}\ z_s(\mu,\varphi)=\min_{\mu\in\Delta(\Delta(A^n))}\ \max_{\varphi\in\Delta(\Delta(A^n))}\ z_s(\mu,\varphi). \tag{8}$$

In the single-expert problem, Nature's randomization 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:

$$z_s(\mu, \varphi) = z_s(P_\mu, \varphi), \qquad \forall \mu, \varphi \in \Delta(\Delta(A^n)). \tag{9}$$

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 reporting the forecast combination $Q = \frac{Q_1+Q_2}{2}$.

The crucial consequence of this asymmetry between Nature's and the expert's randomization is that the values appearing in Eq. 6 and 8 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 see this, consider the setting of Theorem 2 where Nature uses a mixed strategy μ and informs an expert 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. 9 no longer holds. What changed relative to the single-expert model is the availability of data (the informed expert's forecasts) that 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. 9). When randomization is superfluous, as in Eq. 6, the expert is "one step ahead," giving him the advantage.

These observations provide a systematic way to understand why some structural assumptions are critical for the impossibility results. For exam-

ple, why are they inconsistent with repeated sampling, that is so common 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.

5.3 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 game between Nature and a strategic expert as one of hide-and-seek: 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. Testing a deterministic theory is straightforward. The meaning of stochastic truth, on the other hand, is much less obvious. A typical distribution P on outcomes can have infinitely many representations of the form P_μ (or more complicated forms). 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.

6 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.[7] 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

[7] 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.

fundamental incompatibility with other theories in Physics. Yet this theory continues to be maintained because no other theory does better.[8] Economics is full of similar examples. Expected utility theory continue to be the dominant descriptive theory in economic models despite the overwhelming evidence against it. The reason, we suspect, is the lack of convincing alternative.

The classical, frequentist, view attaches probabilities only to events that are subject to repeated identical trials. In the context of testing experts, 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 answer questions like: "If anyone knows the truth, it is expert k." 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.

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

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.

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 (2006a): "Counterfactual Predictions," Northwestern University.

——— (2006b): "Strategic Manipulation of Empirical Tests," Northwestern University.

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