Calibration and Expert Testing, prepared for the Handbook of Game Theory, Volume IV*

Wojciech Olszewski†

January 15, 2014

1 Introduction

Probabilistic estimates of future events have long been playing a significant role in human activity. Probabilistic models are common in science, and are often used in weather forecasting. Many economic markets also rely on probabilistic forecasts, including the forecasts of financial analysts, safety assessors, earthquake locators, traffic flow managers, and sports forecasters.

Probability forecasts can be judged by several criteria (see Murphy and Epstein (1967) for an early study on the topic). Among of most reasonable and objective ones is the criterion of calibration. This criterion has some similarity with the frequency definition of probability, but does not require a background of repeated trials under constant conditions. Dawid (1982) is one of the first theoretical studies of calibration. It shows that if data are generated by a probabilistic model, then forecasts generated by that model are (almost surely) calibrated. Murphy and Winkler (1977) argue that experienced weather forecasters are calibrated.

However, Foster and Vohra (1998) show that one need not know anything about the data-generating process, or be an experienced weather forecaster, in order to be able to produce calibrated forecasts. This is a surprising result. Foster recalls that "this paper took the longest to get published of any I have worked on. I think our first submission was about 1991. Referees simply did not believe the theorem – so they looked for amazingly tiny holes in the proof. When the proof had been compressed from its original 15-20 pages down to about 1, it was finally believed."

The follow-up literature shows that the feature of calibration observed by Foster and Vohra generalizes to a large number of other objective criteria for judging probabilistic models or forecasting. Indeed, suppose that a criterion for judging probabilistic forecasts (which I will call a *test*) has the property that if data

*I am grateful to Nabil Al-Najjar, Dean Foster, Sergiu Hart, Ehud Kalai, Shie Mannor, Alvaro Sandroni, Rann Smorodinsky, Colin Stewart, and Peyton Young for comments on an earlier draft of this survey. I thank the National Science Foundation for research support (CAREER award SES-0644930).

†Department of Economics, Northwestern University, 2001 Sheridan Rd., Evanston IL 60208-2600.

are generated by a probabilistic model, then forecasts generated by that model pass the test. It, then, turns out an agent who knows only the test by which she is going to be judged, but knows nothing about the data-generating process, is often able to pass the test by generating forecasts strategically.

However, this follow-up literature also delivers tests which can be passed by true probabilistic models, but cannot be passed without knowledge of the data-generating process. One can compare the literature surveyed in this paper to non-Bayesian statistics.[1] More specifically, statistics is centered around hypothesis testing. It (implicitly) assumes that the hypotheses being tested were born out of thin air, and were completely unlinked to the hypothesis testing methodology. In particular the hypothesis generating entity had no incentives of its own (or at least they were ignored). The research on testing experts presented in this chapter is all about 'strategic hypothesis testing'. In these papers, we specifically endow the hypothesis generating entity with incentives (and strategies), which is that of passing the 'test'. We rebuilt the notion hypothesis testing, eluding to criteria such as errors of types I and II.

The paper is organized as follows: I first introduce some basic terminology and notation. In Section 3, I present some examples which show how some simple tests can be passed without any knowledge about the data-generating process. Section 4 is entirely devoted to calibration. In Section 5, I continue the exposition of what I call negative, or impossibility results, i.e., the results which say that some tests can be passed without any knowledge about the data-generating process. Positive results, i.e., the results that provide, or prove the existence of, "good" tests are discussed in Section 6. The following three sections are devoted to some results which contrast with the negative results from Section 5, and which have been obtained in slightly different settings. Finally, Section 10 contains some results on philosophy of science and financial markets which are related to, and inspired by the results on testing experts.

2 Terminology and notation

Each period, one out of two possible outcomes 0 or 1 is observed.[2] Define $\Omega = \{0,1\}^{\infty}$ as the set of infinite sequences of outcomes. We will call each $\omega \in \Omega$ a *data set*, or simply, *data*. We will denote the outcome in period t by ω_t, and the *history of outcomes* up to period t by ω^t. That is, $\omega^t = (\omega_1, ..., \omega_{t-1})$ for $t > 1$, and ω^1 means the empty history.

Denote by $\Delta(\Omega)$ the set of all probability measures over Ω. Measures $P \in \Delta(\Omega)$ will sometimes be called stochastic processes. We need a $\sigma-$algebra on which the probability measures are defined. A *cylinder* with base on $(\omega_1, ..., \omega_n)$ is the set of all data sets ω with the first n elements $\omega_1, ..., \omega_n$. We endow Ω with the Borel $\sigma-$algebra, that is, the smallest $\sigma-$algebra that contains all cylinders. We also endow Ω with the product topology, that is, the topology that comprises unions of cylinders.

More generally, for every compact and metrizable space S, denote by $\Delta(S)$ the set of probability

[1] I thank Rann Smorodinsky for suggesting this comparison, wth which I fully agree.

[2] The generalization of the model and all results to any finite set of outcomes is straightforward.

measures on S. We endow $\Delta(S)$ with the *weak*-topology* and with the $\sigma-$algebra of Borel sets (i.e., the smallest $\sigma-$algebra which contains all open sets in weak$^*-$topology). The weak$^*-$topology is defined by the condition that $P^n \to_n P$ if

$$E^{P_n} h \to_n E^P h,$$

for all real-valued and continuous functions h on S, where E is the expected-value operator. In particular, $\Delta(\Delta(\Omega))$ denotes the set of probability measures on $\Delta(\Omega)$. It is well-known that $\Delta(S)$ equipped with the weak*-topology is a compact and metrizable space.

Let $\{0,1\}^t$ denote the Cartesian product of t copies of $\{0,1\}$, and let

$$\Omega^{finite} = \bigcup_{t \geq 0} \{0,1\}^t$$

be the set of all finite histories.[3] Any function

$$f : \Omega^{finite} \longrightarrow \Delta\left(\{0,1\}\right)$$

that maps finite sequences of outcomes into distributions over outcomes will be called a *theory*. A theory takes finite data sets (the outcomes up to a certain period) as inputs, and returns a probabilistic forecast over outcomes for the following period as an output.

It is well-known that every theory f uniquely induces a probability measure $P_f \in \Delta(\Omega)$. More precisely, given a finite history $\omega^k = (\omega_1, ..., \omega_{k-1})$ and an outcome ω_k, let the probability of ω_k conditional on ω^k be denoted by $f(\omega^k)[\omega_k]$. Then, the probability P_f of the cylinder C with base $(\omega_1, ..., \omega_n)$ is equal to the product of probabilities

$$P_f(C) = \prod_{k=1}^{n} f\left(\omega^k\right)[\omega_k].$$

We will often identify theory f with probability measure P_f.

Also, any probability measure $P \in \Delta(\Omega)$ determines a theory f by defining $f(\omega^k)[\omega_k]$ as the probability of ω_k conditional on ω^k. That is, if $P(C(\omega_1, ..., \omega_{k-1})) > 0$, then

$$f(\omega^k)[\omega_k] = \frac{P(C(\omega_1, ..., \omega_k))}{P(C(\omega_1, ..., \omega_{k-1}))},$$

where $C(\omega_1, ..., \omega_k)$ and $C(\omega_1, ..., \omega_{k-1})$ denote the cylinders with base $(\omega_1, ..., \omega_k)$ and $(\omega_1, ..., \omega_{k-1})$, respectively. And $f(\omega^k)[\omega_k]$ is defined in an arbitrary manner if $P(C(\omega_1, ..., \omega_{k-1})) = 0$.[4]

We consider two types of testing. The general definition requires the expert to provide a theory up front, at time 0. But an important class of tests asks for forecasts only along the sequence of observed outcomes. That is, the expert is supposed to provide at the beginning of period $t = 1, 2, ...$ the probability

[3] By convention, $\{0,1\}^0 = \{\varnothing\}$.

[4] This last part of the definition means that there are multiple theories f determined by some probability measures P but, as will become clear shortly, this lack of uniqueness will be irrelevant for our purposes.

of outcome 1 in period t; the expert provides this forecast after observing the outcomes in all previous periods. The expert's forecast of outcome 1 for period t will be denoted by f_t.

Definition 1 *A test is a function*

$$T : \Delta(\Omega) \times \Omega \to \{PASS, FAIL\}.$$

A test is therefore an arbitrary function that takes as an input a theory (more precisely, the probability measure induced by the theory) and the observed sequence of outcomes, and returns as an output a PASS-or-FAIL verdict. In particular, we assume that the verdict is the same for any pair of theories f^1 and f^2 which induce the same probability measure $P_{f^1} = P_{f^2}$ over sequences of outcomes.

We study only measurable tests T. That is, $\{\omega \in \Omega : T(P, s) = PASS\}$ (or equivalently, set $\{\omega \in \Omega : T(P, s) = FAIL\}$) is assumed to be a measurable set for every theory P. The former set will be called the *acceptance set*, and the latter set will be called the *rejection set* for theory P.

We say that the test is *prequential* if the expert is required to give predictions only along the actual sequence of outcomes, i.e., if the verdict of test T depends only on $(f_1, f_2, ...)$. In such a case, we will often write $T(f_1, f_2, ..., \omega)$ instead of $T(P, \omega)$.

We shall now state two basic properties of empirical tests. They are versions of type I and type II errors from statistics. An important conceptual difference (compared to the classic definition of type II error) is that the second property refers to strategic behavior; instead of requiring a false theory to be rejected, we require that an ignorant but strategic expert be rejected.

Definition 2 *Given an $\varepsilon \geq 0$, a test T does not reject the truth with probability $1-\varepsilon$ if, for any $P \in \Delta(\Omega)$,*

$$P(\{\omega \in \Omega : T(P, \omega) = PASS\}) > 1 - \varepsilon.$$

Suppose that there actually is a stochastic process $P \in \Delta(\Omega)$ that generates data. Definition 2 says that a test does not reject the truth if, with high probability, the actual data-generating process P, no matter what that process is, is not rejected. A theory that fails such a test can reliably be viewed as false.

Definition 3 *A test T can be ignorantly passed with probability $1 - \varepsilon$ if there exists a $\xi \in \Delta\Delta(\Omega)$ such that for every sequence of outcomes $\omega \in \Omega$,*

$$\xi(\{P \in \Delta(\Omega) : T(P, \omega) = PASS\}) > 1 - \epsilon.$$

We will call every $\xi \in \Delta\Delta(\Omega)$ a *random generator of theories*. The random generator of theories may depend on test T, but not on any other knowledge, such as knowledge of the actual data-generating process. If a test can be ignorantly passed, we also say that an ignorant expert can pass the test. If a test can be ignorantly passed, then an ignorant but strategic expert can randomly select theories that, with

probability $1-\varepsilon$ (according to the expert's randomization device), will not be rejected, no matter which data set is realized.

In the case of prequential tests, it will sometimes be more convenient to talk about forecasting rules, instead of random generators of theories. A *forecasting rule* specifies, for any history of outcomes ω^t and any history of forecasts $f^t = (f_1, ..., f_{t-1})$, a probability distribution over forecasts f_t. Then, a test can be ignorantly passed if, for every sequence of outcomes $\omega \in \Omega$, the forecasts $(f_1, f_2, ...)$ along ω generated by the rule are such that $T(f_1, f_2, ..., \omega) = PASS$ with high probability.

Suppose that for every random generator of theories $\xi \in \Delta(\Delta(\Omega))$, there exists at least one data set ω such that, with probability greater than ε, the realized theory is rejected on ω. Then, by definition, the test cannot be ignorantly passed with probability $1-\varepsilon$. However, a stronger property may be demanded. A tester may be interested in the existence of data sets such that an ignorant expert fails the test with near certainty (as opposed to probability greater than ε), or may be interested in the existence of a larger number of data sets on which an ignorant expert fails the test.

We will sometimes call a test *good* if it does not reject the truth and cannot be ignorantly passed.

3 Examples

The possibility of ignorantly passing reasonable tests seems quite surprising. Therefore, before presenting more general results, I will use three simple examples to illustrate how this can be achieved.

3.1 Example 1

Consider the following simple test. Let

$$R(f, \omega^{m+1}) = \frac{1}{m}\sum_{t=1}^{m}(f(\omega^t) - \omega_t),$$

where $\omega^t = (\omega_1, ..., \omega_{t-1})$ marks the difference between the average forecast of 1 and the empirical frequency of 1. The test rejects theory f if the average forecast of 1 is not equal to the empirical frequency of 1. That is, theory f is passed on data sets ω such that

$$\lim_m R(f, \omega^{m+1}) = 0.$$

It readily follows from the law of large numbers that the test does not reject the truth. I omit the details of this proof. The test can, however, be ignorantly passed by using the random generator of theories that assigns probability 1 to the single theory f which predicts 1 with certainty in periods in which $R(f, \omega^{m+1}) < 0$, and predicts 0 with certainty in periods in which $R(f, \omega^{m+1}) > 0$. More precisely,

$$f(\omega^{m+1}) = 1 \text{ if } R(f, \omega^{m+1}) < 0,$$
$$f(\omega^{m+1}) = 0 \text{ if } R(f, \omega^{m+1}) > 0,$$
$$f(\omega^{m+1}) = 0.5 \text{ if } R(f, \omega^{m+1}) = 0.$$

The intuition is that when $R(f, \omega^{m+1})$ is negative, the forecast of 1 makes $R(f, \omega^{m+2})$ closer to zero, no matter whether ω_{m+1} is equal to 0 or 1. Similarly, when $R(f, \omega^{m+1})$ is positive, the forecast of 0 makes $R(f, \omega^{m+2})$ closer to zero, no matter whether ω_{m+1} is equal to 0 or 1. I omit the obvious details of the formal proof.

Alternatively, the test can be ignorantly passed by the single theory g which predicts 1 with certainty in periods m such that $\omega_{m-1} = 1$, and predicts 0 with certainty in periods m such that $\omega_{m-1} = 0$; or, more precisely, $f(\omega^m) = \omega_{m-1}$. Notice finally that the ignorant expert must know the test in order to be able pass it ignorantly. One can easily show that no random generator of theories will pass all tests at the same time.

3.2 Example 2

Consider for a moment the setting in which there is only one period, and consider all probability distributions from $\Delta(\{0,1\})$. Any test which does not reject the truth (with probability $1-\varepsilon$) does not reject the truth when the true P assigns equal probabilities to 0 and 1. This implies that

$$T(P,0) = T(P,1) = PASS$$

for any $\varepsilon < 1/2$. Thus, in this case, the test can be ignorantly passed by giving theory P.

Suppose now that there are n periods, where n is such that $1/2^n < \varepsilon$. Denote by $\{0,1\}^n$ the set of all sequences of outcomes $\omega = (\omega_1, ..., \omega_n)$ of length n. Consider test T to be defined as follows: Let m be the lowest number such that $1/2^m < \varepsilon$. For a theory $P \in \Delta(\{0,1\}^n)$, pick any set consisting of 2^{n-m} sequences of outcomes ω such that the probability of this set is the lowest among all sets consisting of 2^{n-m} sequences of outcomes ω. Theory P fails if one of these sequences is observed. By definition, this test passes the truth with probability $1-\varepsilon$.

Since for every P, there exists an ω such that $T(P,\omega) = FAIL$, test T cannot be ignorantly passed by using a degenerated random generator of theories, i.e., by giving a single theory P. Nevertheless, the test can be ignorantly passed. For any history $\omega^{m+1} = (\omega_1, ..., \omega_m)$, take the theory that assigns probability 0 to history ω^{m+1}, and probability $1/(2^m - 1)$ to any other history of the first m outcomes. Randomize uniformly over all such histories or, equivalently, over all such theories.

Given a sequence ω, a theory P that corresponds to some ω^{m+1} fails if the first m outcomes of ω coincide with ω^{m+1}. The probability that the random generator of theories selects such a P is $1/2^m < \varepsilon$.

3.3 Example 3

Consider now the following likelihood ratio (prequential) test. For any theory f, define the alternative theory f^A by letting

$$f^A(\omega^t) = f(\omega^t) + 0.4 \text{ if } f(\omega^t) < 0.5,$$
$$f^A(\omega^t) = f(\omega^t) - 0.4 \text{ if } f(\omega^t) > 0.5,$$
$$f^A(\omega^t) = 0.6 \text{ if } f(\omega^t) = 0.5, \text{ and } t = 1, 3, ...$$
$$f^A(\omega^t) = 0.4 \text{ if } f(\omega^t) = 0.5, \text{ and } t = 2, 4, ...$$

Define the likelihood of outcome ω_t according to theory f by

$$l(\omega_t) = f(\omega^t) \text{ if } \omega^t = 1, \text{ and } l(\omega_t) = 1 - f(\omega^t) \text{ if } \omega^t = 0,$$

and let $l^A(\omega_t)$ be defined similarly. For any sequence of outcomes $\omega^{t+1} = (\omega_1, ..., \omega_t)$, let

$$L(\omega^{t+1}) = \frac{l(\omega_1) \cdot ... \cdot l(\omega_t)}{l^A(\omega_1) \cdot ... \cdot l^A(\omega_t)}$$

be the likelihood ratio of ω^{t+1} according to f compared to the alternative theory.

Finally, define test T by letting $T(f_1, f_2, ..., \omega) = PASS$ if

$$\lim_t L(\omega^{t+1}) = \infty, \tag{1}$$

and $T(f_1, f_2, ..., \omega) = FAIL$ otherwise. That is, test T passes theory f if the observed sequence of outcomes is infinitely more likely according to theory f than according to theory f^A.

It readily follows from the law of large numbers that the test does not reject the truth. I omit the details of this proof. The test can be ignorantly passed by the forecasting rule that predicts $f_t = 0.4$ with probability $1/2$ and $f_t = 0.6$ with probability $1/2$, independent of the history of outcomes up to period t.

The intuition is that if the observed outcome was predicted by the expert as more likely (i.e., $\omega_t = 0$ and $f_t = 0.4$, or $\omega_t = 1$ and $f_t = 0.6$), then the ratio $l(\omega_t)/l^A(\omega_t)$ is $0.6/0.2 = 3$, while if the observed outcome was predicted by the expert as less likely (i.e., $\omega_t = 0$ and $f_t = 0.6$, or $\omega_t = 1$ and $f_t = 0.4$), then the ratio $l(\omega_t)/l^A(\omega_t)$ is only $0.4/0.8 = 1/2$. This gives the expert's theory an advantage over the alternative theory. In order to satisfy condition (1), it suffices that the expert predicts with frequency $1/2$ the outcome which is later observed as more likely.

The fact that the likelihood test can be ignorantly passed follows from the law of large numbers. I again omit the details of the formal proof.

4 Calibration

4.1 Definition and result

The existing literature contains several similar definitions of calibration; they are not all equivalent. In this survey, calibration is defined as follows: Just before time t, after all previous outcomes have been

observed, a forecast f_t is made of the probability that $\omega_t = 1$. It is assumed that this forecast takes on values that are the midpoints of one of the intervals: $[0, 1/m], [1/m, 2/m], ..., [(m-1)/m, 1]$. That is,

$$f_t = M_i = \frac{2i-1}{2m},\ i = 1, ..., m.$$

Let

$$I_{f_t=M_i} = 1 \text{ if } f_t = M_i, \text{ and } I_{f_t=M_i} = 0 \text{ if } f(\omega^t) \neq M_i$$

be the indicator function of the set $\{\omega^t : f_t = M_i\}$. The empirical frequency ρ_i^T of outcome 1, where $i = 1, ..., m$, is defined as

$$\frac{\sum_{t=1}^{T} \omega_t I_{f_t=M_i}}{\sum_{t=1}^{T} I_{f_t=M_i}},$$

if $I_{f_t=M_i} = 1$ for some t; and

$$\rho_i^T = \frac{2i-1}{2m}$$

if $I_{f_t=M_i} = 0$ for all t, where ω_t is the outcome observed in period t. The empirical frequency ρ_i^T is the frequency with which outcome 1 is observed in those periods $t < T$ for which the forecast is $f_t = M_i$.

Finally, let

$$\overline{I}^T_{f_t=M_i} = \frac{1}{T} \cdot \sum_{t=1}^{T} I_{f_t=M_i}$$

be the frequency of forecast M_i. A sequence of forecasts $(f_t)_{t=1}^{\infty}$ is $(1/m)-$*calibrated* if

$$\limsup_T \left|\rho_i^T - M_i\right| \leq \frac{1}{2m}$$

for every $i = 1, ..., m$ such that

$$\limsup_T \overline{I}^T_{f_t=M_i} > 0.$$

That is, if forecast $f_t = M_i$ is being made in a positive fraction of periods, then the limit empirical frequency ρ_i^T must be as close to M_i as possible, given the assumption that the forecasts must have the form $f_t = M_i$ for some $i = 1, ..., m$. If number m is sufficiently large, then we say that the forecasts are approximately calibrated.

Proposition 1 *(Foster and Vohra (1998)) For every m, there exists a forecasting rule ξ such that for every ω, the sequence of forecasts $(f_t)_{t=1}^{\infty}$ generated by ξ along the sequence of outcomes ω is almost surely $(1/m)-$calibrated.*

4.2 Calibrated forecasting rule

A forecasting rule with the required property is fairly easy to define. Given a theory f, and a history ω^T, let

$$\overline{d}^i = (\frac{i-1}{m} - \rho_i^T) \cdot \overline{I}^T_{f(\omega^t)=M_i}$$

and

$$\overline{e}^i = (\rho_i^T - \frac{i}{m}) \cdot \overline{I}^T_{f(\omega^t)=M_i}.$$

We define ξ by specifying a probability distribution over forecasts at each ω^T as follows:

(1) If there exists an i such that $\rho_i^T \in [(i-1)/m, i/m]$ (or, equivalently, $\overline{d}^i \leq 0$ and $\overline{e}^i \leq 0$), then $f(\omega^{T+1}) = M_i$ for any i with this property.

(2) Otherwise, there exists an i such that $\overline{d}^i > 0$ and $\overline{e}^{i-1} > 0$. Then $f(\omega^{T+1}) = M_i$ with probability

$$\frac{\overline{e}^{i-1}}{\overline{d}^i + \overline{e}^{i-1}},$$

and $f(\omega^{T+1}) = M_{i-1}$ with probability

$$\frac{\overline{d}^i}{\overline{d}^i + \overline{e}^{i-1}}.$$

A simple inductive argument shows that if there is no i such that $\rho_i^T \in [(i-1)/m, i/m]$, then there exists an i such that $\overline{d}^i > 0$ and $\overline{e}^{i-1} > 0$. Indeed, from the definition of $\overline{d}^1$, we have that $\overline{d}^1 \leq 0$. So if $\overline{e}^1 \leq 0$, condition (1) is satisfied by $i = 1$. If $\overline{e}^1 > 0$ and $\overline{d}^2 > 0$, then condition (2) is satisfied by $i = 2$. Otherwise, $\overline{d}^2 \leq 0$, and one can apply the previous argument again. Finally, if $\overline{d}^m \leq 0$, then by the definition of $\overline{e}^m$, we have that $\overline{e}^m \leq 0$, and so condition (1) is satisfied by $i = m$.

This forecasting rule, which achieves a high calibration score, can be best understood in the case of $m = 2$. If the current empirical frequency of outcome 1 over the periods in which the forecast was 1/4 happens to belong to $[0, 1/2]$, then predict that the current-period outcome will also belong to $[0, 1/2]$, that is, predict 1/4. Similarly, predict 3/4 if the current empirical frequency of outcome 1 over the periods in which the forecast was 3/4 happens to belong to $[1/2, 1]$. Choose either of the two forecasts if both empirical frequencies belong to the appropriate intervals.

Otherwise, the former empirical frequency is higher than 1/2, and the latter empirical frequency is lower than 1/2. In this case, randomize over forecasts 1/4 and 3/4. Assign to each of the two forecasts a probability that is inversely proportional to the distance of the empirical frequency to the appropriate interval, namely, $\overline{e}^1/(\overline{d}^2 + \overline{e}^1)$ and $\overline{d}^2/(\overline{d}^2 + \overline{e}^1)$, respectively.

4.3 Sketch of proof

Although the forecasting rule is simple, the proof that it actually achieves a high calibration score (that is, the proof of Proposition 1) is not that simple. I will sketch the proof which comes from Foster (1999) in the case of $m = 2$. This proof is a simplification of the general by Sergiu Hart and Andreu Mas-Colell proof for the case in which in every period only two outcomes are possible. In order to prove Proposition 1, we need to recall the concept of approachability and the celebrated theorem from Blackwell (1956).

Consider a two-person, zero-sum game in which each player takes actions from a finite set. Each player's payoff is an $L-$dimensional vector. For our purposes, it is convenient to denote the actions by $i = 1, 2, ..., m$ and $x \in X$, and the payoff of the first player, who will be called player I, by $c(i, x)$. This game is played repeatedly over time. Let C be a closed and convex subset of R^L. We call C *approachable* by player I if there exists a repeated-game strategy of player I which guarantees that the average payoff of player I almost surely converges to set C, regardless of the actions of player I's opponent, as the number of repetitions converges to infinity.

Theorem 1 *(Blackwell (1956)) A set C is approachable if and only if for all $a \in R^L$, there exists a vector $w \in \Delta(\{1, ..., m\})$ such that for all $x \in X$,*

$$\sum_{i=1}^{m} w_i (c(i, x) - b)^T \cdot (a - b) \leq 0, \tag{2}$$

where b is the closest point to a in C.[5]

Moreover, it follows from the proof that C is approachable if condition (2) is satisfied for all vectors a that have the form of average payoffs of player I up to time $T = 1, 2, ...$ Then, w_i is the probability of taking action i in period T, when the average payoff up to period T is a. I will sketch the proof of Blackwell's theorem at the very end of this section.

For our purposes, $L = 2$ (since we are sketching the proof for $m = 2$), and

$$C = \{z = (z_1, z_2) \in R^2 : z_1, z_2 \leq 0\}. \tag{3}$$

Player I is the expert, action $i = 1, 2$ represents forecast M_i. One can think of player I's opponent as nature, and the set of outcomes $X = \{0, 1\}$ as nature's actions. The payoff vector $c(i, x)$ is defined as

$$c(i, x) = (d^2(i, x), e^1(i, x)), \tag{4}$$

where

$$d^2(i, x) = \begin{cases} 0 \text{ if } i = 1 \\ 1/2 - x \text{ if } i = 2 \end{cases}$$

and

$$e^1(i, x) = \begin{cases} x - 1/2 \text{ if } i = 1 \\ 0 \text{ if } i = 2 \end{cases}.$$

In order to prove Proposition 1, we need to show that condition (2) is satisfied. Notice that given $a = (a_1, a_2)$,

$$b = (b_1, b_2) = (\min\{0, a_1\}, \min\{0, a_2\})$$

and

$$a - b = (\max\{0, a_1\}, \max\{0, a_2\});$$

[5] By $(c(i, x) - b)^T \cdot (a - b)$ we denoted the inner product of vectors $(c(i, x) - b)^T$ and $(a - b)$, and the distance between two points in R^L is measured in the standard manner.

since $\max\{0, a_i\} \cdot \min\{0, a_i\} = 0$,

$$\sum_{i=1}^{2} w_i (c(i,x) - b)^T \cdot (a - b) = \sum_{i=1}^{2} w_i c(i,x)^T \cdot (a - b) =$$

$$= \sum_{i=1}^{2} w_i [d^2(i,x) \max\{0, a_1\} + e^1(i,x) \max\{0, a_2\}].$$

By the definition of $d^2(i,x)$ and $e^1(i,x)$, this is equal to

$$w_1(x - 1/2) \max\{0, a_2\} + w_2(1/2 - x) \max\{0, a_1\}$$

This expression is equal to zero if we take $w_1 = 1$ and $a_2 = \overline{e}^1(1,x) \leq 0$, or if we take $w_2 = 1$ and $a_1 = \overline{d}^2(2,x) \leq 0$. Otherwise, $\overline{d}^2(2,x) > 0$ and $\overline{e}^1(1,x) > 0$. Then, since we take $w_1 = \overline{e}^1(1,x) / \left(\overline{d}^2(2,x) + \overline{e}^1(1,x)\right)$ and $w_2 = \overline{d}^2(2,x) / \left(\overline{d}^2(2,x) + \overline{e}^1(1,x)\right)$, the expression becomes

$$\frac{\overline{d}^2(2,x)}{\overline{d}^2(2,x) + \overline{e}^1(1,x)} (1 - x) \overline{e}^1(1,x) + \frac{\overline{e}^1(1,x)}{\overline{d}^2(2,x) + \overline{e}^1(1,x)} (x - 1) \overline{d}^2(2,x) = 0.$$

This completes the proof that condition (2) is satisfied.

4.4 Sketch of proof of Blackwell's theorem

The idea of the proof behind Blackwell's theorem can be explained as follows: Denote player I's average payoff at time t by A_t, the expected average payoff at time $t+1$ contingent on A_t by $E[A_{t+1} \mid A_t]$, and the expected average payoff at time $t+1$ from the perspective of period 0 by $E[A_{t+1}]$. Then,

$$E[A_{t+1} \mid A_t] = \frac{t}{t+1} A_t + \frac{1}{t+1} E[c(i,x) \mid A_t].$$

Since the strategy of player I has the property that the inner product of $A_t - B_t$ and $E[c(i,x) \mid A_t] - B_t$ is nonpositive (B_t stands for the point in C that is closest to A_t), it follows that $E[A_{t+1} \mid A_t]$ is closer to set C than A_t is (see Figure 1).

Moreover, $E[A_{t+1}]$ converges to C, because the norm of the second component of $E[A_{t+1} \mid A_t]$ is at least of order $1/t$ of the norm of the first component. Together with the fact that the inner product of the two components is nonpositive, this implies that the distance between $E[A_{t+1}]$ and C shrinks in period t at least by order $1/t$. But if $E[A_{t+1}]$ did not converge to C, this would mean that the series

$$\sum_t \frac{1}{t}$$

was convergent.

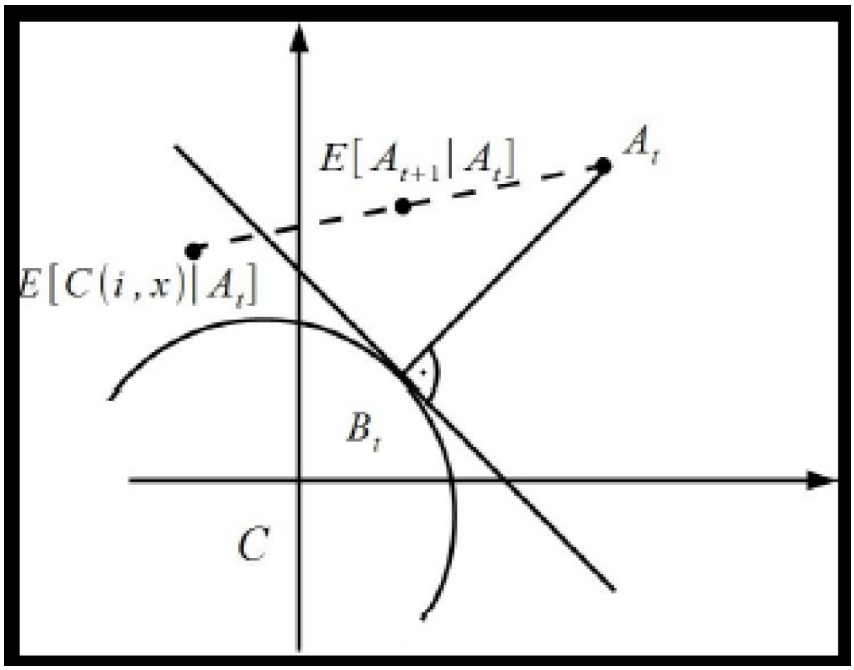

Figure 1

Now, some form of the law of large numbers (more precisely, the strong law of large numbers for dependent variables) implies that A_{t+1} converges to C almost surely.

5 Negative results

Early follow-up papers generalize the Foster and Vohra (1998) result to "other forms" of calibration. Other papers provide simpler proofs of their result. More specifically, a *history-based checking rule* is an arbitrary function of finite sequences of outcomes (histories) to the set {active, inactive}. Given a history-based checking rule and a theory, a *forecast-based checking rule* is active if the history-based checking rule is active and the forecast takes a value from a given set $D \subset [0,1]$. The calibration score assigned to a theory by a checking rule (either history- or forecast-based) is the difference between the frequency of an outcome and the average probability assigned to this outcome by the forecasts of the theory, where the averages are taken across the periods in which the checking rule was active. We follow here the terminology introduced in Sandroni et al. (1999).

According to this terminology, Foster and Vohra (1998) demonstrate the existence of a forecasting rule that (almost surely) calibrates the forecast-based checking rules associated with the always active history-based rule and sets $D_k = [k/m, (k+1)/m]$ (where $k = 0, ..., m-1$). The concept of checking rules other than calibration (i.e., other notions of calibration) was introduced in Kalai et al. (1999), which also demonstrates equivalence between notions of calibration and merging. Lehrer (2001) shows that there exists a forecasting rule which simultaneously calibrates any countable number of history-based

checking rules; Sandroni et al. (1999) generalizes this result to the forecast-based checking rules associated with a countable number of history-based checking rules and a countable number of sets D. Lehrer also shows that for any probability distribution over history-based checking rules, there exists a forecasting rule which simultaneously calibrates almost all these rules. All these results allow for any finite set of possible outcomes, not only 0 and 1.

Sergiu Hart first suggested a simpler proof of Foster and Vohra's result based on the minmax theorem (see the discussion in Foster and Vohra (1998) and Sandroni (2003)). A constructive version of Hart's proof has been derived independently by Fudenberg and Levine (1999). Foster (1999) contains the simple and elegant proof that has been discussed in the previous section. (See also Foster and Vohra (1997) and Hart and Mas-Colell (2000).)[6] The last two authors have suggested using Blackwell's theorem. Foster and Vohra's original argument is close to the proof based on Blackwell's theorem – but uses a direct potential function instead of Blackwell's theorem. Sandroni (2003) first observed that the Foster and Vohra result generalizes well beyond calibration and scoring rules. (See also Vovk and Shafer, 2005, and Vovk 2007.) I discuss Sandroni's result, and the generalizations of thereof, in the following section.

5.1 Generalizations of Foster and Vohra's result

Theorem 2 *(Fan (1953)) Let X be a convex subset of a compact, Hausdorff, linear topological space, and Y be a convex subset of a linear space (not necessarily topologized). Let f be a real-valued function on $X \times Y$ such that for every $y \in Y$, $f(x, y)$ is lower semi-continuous on X. If f is convex on X and concave on Y, then*

$$min_{x \in X} sup_{y \in Y} f(x, y) = sup_{y \in Y} min_{x \in X} f(x, y). \tag{5}$$

Theorem 2 is illustrated in Figure 2, where $X = Y = [0, 1]$, and f is a linear function of x and a linear function of y. The nontrivial part of the theorem says that the right-hand side can be as large as the left-hand side. Suppose that the left-hand side of (5) is "large." This means that for every x there is a y such that $f(x, y)$ is large. In Figure 2, $y = 1$ for $x = 0$ and $y = 0$ for $x = 1$. The linearity and continuity of $f(x, y)$ imply that there is a value of x (between 0 and 1) such that $f(x, y)$ is a constant function of y for this x, and is depicted in bold on the graph of function f. This constant must be large, since there must exist a y for this x such that $f(x, y)$ is large. However, by analogous arguments, there also exists a value of y such that $f(x, y)$ is a constant function of x for this y. This is also depicted in bold on the graph of function f. And this constant is large, because the two bold lines on the graph intersect. Thus, the right-hand side of (5) must also be large.

[6] Foster's proof allows for only two outcomes: 0 and 1. A simple proof which allows for any finite number of outcomes is provided in Mannor and Stoltz (2010).

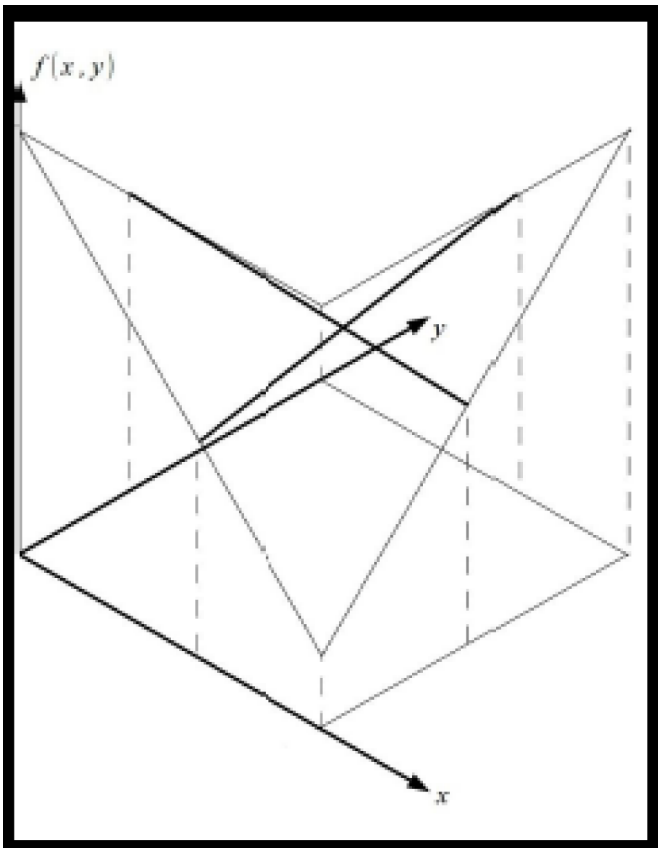

Figure 2

I have stated the minmax theorem in its original form. However, I will not define Hausdorff spaces or lower semi-continuity. It is enough to know that all spaces considered in this survey will be metrizable, and therefore Hausdorff, and all functions will be continuous.

We will call a test T *finite* if for every $P \in \Delta(\Omega)$, there exists an n such that for any $\omega(1), \omega(2) \in \Omega$ such that $\omega^n(1) = \omega^n(2)$, we have

$$T(P, \omega(1)) = T(P, \omega(2)).$$

That is, the verdict of a finite test T depends only on the a finite number n of observed outcomes. Notice, however, that I allow this finite number to depend on theory P.

Fix $\varepsilon \in [0, 1]$ and $\delta \in (0, 1 - \varepsilon]$.

Proposition 2 *(Sandroni (2003), Olszewski and Sandroni (2008))*[7] *Let T be a finite test that does not reject the truth with probability $1 - \varepsilon$. Then, the test T can be ignorantly passed with probability $1 - \varepsilon - \delta$.*

Proof. To provide the intuition for Proposition 2 consider the following zero-sum game between nature and the expert. Nature chooses a probability measure $P \in \Delta(\Omega)$. The expert chooses a random generator

[7] Sandroni (2003) proved this result for tests which are finite in a slightly stronger sense, namely, he assumes that the number n in the definition of a finite test is the same for all theories P.

of theories ξ. The expert's payoff is

$$E^{\xi}E^{P}T(Q,\omega),$$

where $PASS = 1$ and $FAIL = 0$, and E^{ξ} and E^{P} are the expectation operators associated with ξ and P, respectively. By assumption, test T does not reject the truth with probability $1-\varepsilon$. Thus, for every strategy P of nature, there is a strategy ξ_P for the expert (that assigns probability one to P) such that the expert's payoff is

$$E^{\xi_P}E^{P}T(Q,\omega) = P\{\omega \in \Omega : T(P,\omega) = 1\} \geq 1-\varepsilon.$$

Hence, if the zero-sum game has value, then there is a strategy ξ_T for the expert that ensures a payoff arbitrarily close to $1-\varepsilon$, no matter which strategy nature chooses. In particular, nature can use P_ω, selecting a single sequence of outcomes ω with certainty. Therefore, for all $\omega \in \Omega$,

$$E^{\xi_T}E^{P_\omega}T(Q,\omega) = \zeta_T\{Q \in \Delta(\Omega) : T(Q,\omega) = PASS\} \geq 1-\epsilon-\delta.$$

Fan's minmax theorem guarantees that the zero-sum game between nature and the expert has value. More precisely, let X be $\Delta(\Omega)$, and let Y be the subset of $\Delta(\Delta(\Omega))$ consisting of all random generators of theories with finite support. That is, an element ξ of Y can be described by a finite sequence of probability measures $Q^1,...,Q^n$ and positive weights $\pi^1,...,\pi^n$ that add up to one such that ξ selects Q^i with probability π^i,$i = 1,...,n$. Let function $f : X \times Y \to R$ be defined by

$$f(P,\xi) = E^{\xi}E^{P}T(Q,\omega) = \sum_{i=1}^{n}\pi^i \int T(Q^i,\omega)dP(\omega).$$

All the conditions of the minmax theorem are satisfied. In particular, function f is continuous in P and linear, as the sum of continuous and linear functions. The continuity of functions of the form

$$E^{P}T(Q^i,\cdot) = \int T(Q^i,\omega)dP(\omega)$$

follows immediately from the assumption that test T is finite and from the definition of weak*-topology; this guarantees that $T(Q^i,\omega)$ is a continuous function of ω. ■

5.2 Prequential principle

I conclude this section with two recent results, which show that if there is a good test, it must make use of counterfactual forecasts, which cannot be verified by any observed data.

We will say that a test rejects theories in finite time if sets $\{\omega \in \Omega : T(P,s) = FAIL\}$ are unions of cylinders.

For every theory, such a test specifies a collection of finite sequences of outcomes, which sequences contradict the theory according to the test; it therefore fails the theory if one of these sequences is observed.

Two theories f^1 and f^2 are equivalent up to period m if

$$f^1\left(\omega^{k+1}\right) = f^2\left(\omega^{k+1}\right)$$

for any $\omega^{k+1} = (\omega_1, ..., \omega_k)$ such that $k < m$. Two theories are therefore equivalent up to period m if they make the same predictions for periods $1, ..., m$.

Definition 4 *A test T is future-independent if, for any pair of theories f^1 and f^2 that are equivalent up to period m, and for any sequence of outcomes $\omega^{m+1} = (\omega_1, ..., \omega_m)$, theory f^1 is rejected at ω^{m+1} if and only if theory f^2 is rejected at ω^{m+1}.*

A test is future-independent if, whenever a theory f^1 is rejected in period m, another theory f^2, which makes exactly the same predictions as f^1 up to period m, must also be rejected in period m. In other words, if a finite sequence of outcomes contradicts a theory, then it also contradicts any theory equivalent to it.

Proposition 3 *(Olszewski and Sandroni (2008)) Every future-independent test which rejects theories in finite time, and which, with probability $1 - \varepsilon$, does not reject the truth, can be ignorantly passed with probability $1 - \varepsilon - \delta$.*

Recall that a test is prequential if it requires the expert to give predictions only along the actual sequence of outcomes.

Proposition 4 *(Shmaya (2008)) Every prequential test T that, with probability $1 - \varepsilon$, does not reject the truth, can be ignorantly passed with probability $1 - \varepsilon - \delta$.*

Propositions 3 and 4 are independent in that neither of them implies the other. There exist future-independent tests whose verdicts depend on counterfactual, "off-equilibrium" predictions. There also exist tests which require the expert to give predictions only along the actual sequence of outcomes, and which may not reject theories in finite time. Tests that reject theories in finite time have the property that sets $\{\omega \in \Omega : T(P, \omega) = PASS\}$ are closed in Ω. Shmaya's result allows for tests such that these sets are Borel, but may not be closed.

5.3 Interpretations

There are two interpretations of the results on testing experts. One is literal and involves informed versus ignorant experts. Informed experts know precisely the probability distribution P that generates data, and ignorant experts are completely ignorant, without even any prior over probability distributions. Although, this language is very convenient, and I am using it throughout this survey, I would argue that it should not be taken too literally.

The literal interpretation faces a number of conceptual problems. For example, what does it mean to know the probabilities of future events? But even if the concept of probability is taken as given, it is unclear whether the existence of a random generator of theories which satisfies Definition 3 really helps an ignorant but strategic expert.

To see the point, assume, as the literature often does, that an ignorant expert must make the decision whether to provide her forecasts. These forecasts will be tested. The expert receives a positive utility u from providing the forecasts, but if she fails the test, she also receives a disutility $-d$, such that $u - d < 0$. That is, the expert's utility depends on the verdict of the test, and thus on her forecasts and the observed sequence of outcomes. The ignorant expert does not know which sequence of outcomes will be observed at the time the forecasts are provided; moreover, the expert is completely ignorant, which means that she does not even have any prior over the sequences of outcomes. In other words, the expert faces Knightian uncertainty (also known as modern uncertainty or ambiguity). Suppose that the expert is most uncertainty-averse in the sense of Gilboa and Schmeidler (1989); this means that she evaluates prospects according to the worst possible scenario, i.e., the sequence of observed outcomes which gives her the lowest utility.

The existence of a random generator of theories ξ that makes the probability of passing the test high, even without any knowledge regarding the data-generating process, seemingly makes the option of providing forecasts more attractive. Put another way, if the expert forecasts according to ξ, then for every possible sequence of observed outcomes, she will end up, according to ξ, with utility $u - d$ with a low probability, and with utility u with a high probability.

However, this argument is no different from Raiffa's (1961) critique of the concept of uncertainty. Raiffa claimed that an appropriate randomization can remove uncertainty and replace it with common risk, and therefore the presence of uncertainty should have no more impact on a decision maker's utility than the presence of common risk. Subsequent studies on decision theory tend to disagree with Raiffa's critique. Intuitively, the reason for this disagreement is that the expert can randomize only before the uncertainty regarding sequence ω is resolved; as a result, once lottery ξ is resolved and a probability distribution P is selected, the decision maker faces uncertainty again. If, given the P selected, she again evaluates prospects according to the sequence of observed outcomes that gives her the lowest utility, she will typically end up with utility $-d$, because nontrivial tests fail every P on some ω.

Of course, some tests (for example, the test from Example 1) can be passed without randomizing over theories, and passing of some other tests requires "little" randomization. However, the literal interpretation of the negative results on testing experts seems to require that random generators of theories have some properties in addition to the property from Definition 3, and therefore future is necessary.

In my opinion, this literature is about the philosophy of science, or more precisely, about probabilistic models; this is the alternative to the literal interpretation. One may disagree about the way we should interpret the concept of probability, but probabilistic models are nevertheless commonly used in scientific practice. So, if we want to test them, which is also a common scientific practice, we would do better to have tests that do not reject them when they are correct. That is, if we can generate data according to a probabilistic model, the test should not reject the model with high probability. For example, if we agree that we have a fair coin, then flipping it repeatedly should generate data sequences such that the i.i.d. fifty-fifty model will pass the test most of the time. This is the way I interpret the condition that the test

passes the truth.

In my view, the major drawback in having a test that can be ignorantly passed is not that such a test cannot differentiate between informed and ignorant decision makers, but rather that the test is vulnerable to the kind of actions of malicious agents that computer scientists study. Actually, I believe that scientific testing methods should be designed so as to exclude any test that can be ignorantly passed, i.e., no test should be passable by a person with malicious intent, who is ignorant about a particular forecasting problem but who is otherwise smart.

6 Positive results

6.1 Category tests

After the initial wave of negative results, described in Section 5, few new results appear for a short period. The revitalizing paper was Dekel and Feinberg (2006), which offered an intuitive idea of constructing a good test. Their starting point was a well-known result from measure and category theory. To get to this result, consider first the following definition:

Definition 5 *A subset of a compact metric space is topologically large (i.e., residual) if it contains the intersection of a countable family of open and dense sets. A subset is topologically small (i.e., first-Baire category) if it is contained in a set whose complement is topologically large.*

I refer the reader to Oxtoby (1980) for a more complete exposition of the basic concepts of category.

Theorem 3 *For every measure P defined on the $\sigma-$algebra of Borel subsets of a compact metric space, there exists a topologically large subset G of the metric space such that*

$$P(G) = 0.$$

Since set G contains the intersection of a countable family of open and dense sets, it follows that for every $\varepsilon > 0$, there also exists an open and dense subset U such that $P(U) \leq \varepsilon$.

Dekel and Feinberg assume that the expert must give a theory at time 0, and suggest a class of tests that they call *category tests*, which pass any theory only on a topologically small set. By Theorem 3, there exist tests in this class which pass the truth with probability 1, and the intuition suggests that the ignorant expert should find passing a category test difficult. For strategic reasons, the expert picks a topologically small set on which she wants to pass. So, for any data, the expert must pick with high probability a topologically small set containing that data, without knowing anything about the data-generating process.

To support this intuition, Dekel and Feinberg show that for every topologically small set S, the set of theories P such that $P(S) > 0$ is topologically small. Their intuition is, however, not entirely correct, since Olszewski and Sandroni (2011) exhibits a category test that can be ignorantly passed. In spite of that,

Dekel and Feinberg provide an example of good category test. To construct such a test, they use a *Lusin set*, which is a subset of Ω with certain "exotic" properties. The existence of Lusin sets is independent of the usual axioms of set theory, and for this reason, I omit the details of Dekel and Feinberg's construction. I will provide another good category test later, but it will be useful to first provide a simpler good test with somewhat weaker properties.

6.2 A simple example of a good test

This simple example of a good test is provided in Olszewski and Sandroni (2011): Consider any sequence of pairwise-disjoint, nonempty sets $C_n \subset \Omega$. For concreteness, let set C_n be cylinders $C(\omega^{n+1})$, $n = 0, 1, ...$, where ω^{n+1} is the finite history in which outcome 1 was observed in periods $1, ..., n-1$, and outcome 0 is observed in period n. For any theory P, consider the sets of the form

$$\overline{C}_n = \bigcup_{k=n}^{\infty} C_k.$$

There exist a number n such that $P(\overline{C}_n) < \varepsilon$, because

$$\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} C_k = \varnothing.$$

Define $R(P)$ as $\overline{C}_n$ for the lowest number n such that $P(\overline{C}_n) < \varepsilon$.

The test requires the expert to provide a theory P up front (at time 0), and rejects the theory when data $\omega \in R(P)$ are observed. In other words, theory P is rejected if outcome 0 is observed after observing a sufficiently long sequence of outcomes 1.

By definition, the test rejects the truth with probability no higher than ε. It turns out that the test cannot be ignorantly passed. Indeed, for any random generator of theories $\xi \in \Delta(\Delta(\Omega))$, consider sets

$$\mathbf{Q}_n = \left\{ Q \in \Delta(\Omega) : R(Q) := \bigcup_{k=n}^{\infty} C_k \right\}.$$

Notice that the family of sets $\mathbf{Q}_n$ is a partition of $\Delta(\Omega)$. It follows that

$$\sum_{n=1}^{m} \xi(\mathbf{Q}_n) \geq 1 - \varepsilon$$

when m is sufficiently large. And thus,

$$\xi\left(\{Q \in \Delta(\Omega) : T(Q, \omega) = PASS\}\right) \leq \varepsilon$$

for any data set

$$\omega \in \bigcup_{k=m}^{\infty} C_k;$$

we denote this set of data sets ω by R_m. That is, by generating theories according to any ξ, the ignorant expert (like the informed expert) passes the test with probability ε or lower, if the tester observes 0 after a

sufficiently long sequence of 1's. However, unlike the informed expert, the ignorant expert does not know how likely it is that the tester will observe data on which she will fail the test.

This simple test has a number of drawbacks. One of them is that although the ignorant expert cannot make sure that she will pass the test on all data sets, she has strategies enabling her to pass the test on "almost all data sets." More precisely, there exists a sequence of random generators of theories $(\xi_m)_{m=1}^{\infty}$ such that by generating theories according to ξ_m, the ignorant expert fails only on set R_m, and $(R_m)_{m=1}^{\infty}$ is a descending sequence of sets with empty intersection. Indeed, by reporting a theory $Q \in \mathbf{Q}_m$, the expert fails only on set R_m.

6.3 Other good tests

If we combine the simple idea for constructing a good test from the previous section, with the Dekel and Feinberg (2006) concept of category, we obtain tests with much stronger properties; in particular, we obtain tests that cannot be ignorantly passed on almost all data sets.

Proposition 5 *(Olszewski and Sandroni (2009c)) (a) For every $\varepsilon > 0$, there exists a test T that passes the truth with probability $1-\varepsilon$, and cannot be ignorantly passed. Moreover, for every random generator of theories $\xi \in \Delta(\Delta(\Omega))$, there is an open and dense set $U \subset \Omega$ such that*

$$\xi\left(\{Q \in \Delta(\Omega) : T(Q,\omega) = PASS\}\right) \leq \varepsilon, \ \forall\ \omega \in U.$$

(b) There also exists a test T that passes the truth with probability 1 such that for every random generator of theories $\xi \in \Delta(\Delta(\Omega))$, there is a topologically large set $G \subset \Omega$ such that

$$\xi\left(\{Q \in \Delta(\Omega) : T(Q,\omega) = PASS\}\right) = 0, \ \forall\ \omega \in G.$$

Proof. I will prove part (a) only. The proof of part (b) is analogous, although slightly more involved. Let $D = \{\omega(1), \omega(2), ...\} \subset \Omega$ be any countable and dense subset of Ω. For concreteness, one can assume that D consists of all data sets such that, from some period on, only outcome 1 is observed.

For any theory P, take the lowest n such that

$$P\left(C(\omega^{n+1}(k)) - \{\omega(k)\}\right) < \frac{\varepsilon}{2^k}. \tag{6}$$

That is, I consider cylinders with base on $\omega^{n+1}(k)$,[8] from which the point $\omega(k)$ has been removed, and take such a cylinder whose measure P is appropriately small. There exists an n with property (6), because

$$\bigcap_{n=0}^{\infty} \left[C(\omega^{n+1}(k)) - \{\omega^k\}\right] = \varnothing.$$

Denote by $C_k(P)$ the cylinder $C(\omega^{n+1}(k))$ with property (6).

[8] Recall that history $\omega^{n+1}(k)$ consists of the first n outcomes of $\omega(k)$.

Let

$$R(P) = \bigcup_{k=1}^{\infty} [C_k(P) - \{\omega(k)\}].$$

The test T rejects theory P when $\omega \in R(P)$ is observed.

The test rejects the truth with probability no higher than ε, because

$$P(R(P)) \leq \sum_{k=1}^{\infty} P(C_k(P) - \{\omega(k)\}) < \sum_{k=1}^{\infty} \frac{\varepsilon}{2^k} = \varepsilon.$$

Now, take any random generator of theories $\xi \in \Delta(\Delta(\Omega))$, and consider sets

$$\mathbf{Q}_{k,n} = \left\{Q \in \Delta(\Omega) : C_k(Q) = C(\omega^{n+1}(k))\right\}.$$

Notice that the family of sets $\mathbf{Q}_{k,n}$, $n = 1, 2, ...$, is a partition of $\Delta(\Omega)$. It follows that

$$\sum_{n=m(k)+1}^{\infty} \xi(\mathbf{Q}_{k,n}) \leq \varepsilon$$

for any sufficiently large $m(k)$.

Let

$$U = \bigcup_{k=1}^{\infty} \left[C(\omega^{m(k)+1}(k))) - \{\omega(k)\}\right].$$

Set U is an open and dense in Ω. Take any $\omega \in U$. Then, $\omega \in C(\omega^{m(k)+1}(k)) - \{\omega(k)\}$ for some k, which means that

$$\{Q \in \Delta(\Omega) : T(Q, \omega) = PASS\} \subset \bigcup_{n=m(k)+1}^{\infty} \mathbf{Q}_{k,n},$$

because any $Q \in \mathbf{Q}_{k,n}$ for $n \leq m(k)$ is rejected on $C(\omega^{n+1}(k)) - \{\omega(k)\} \supset C(\omega^{m(k)+1}(k)) - \{\omega^k\}$. It follows that

$$\{Q \in \Delta(\Omega) : T(Q, \omega) = PASS\} \leq \varepsilon.$$

■

6.4 Good "prequential" tests

Proposition 4 assumes that the expert must give a theory at time 0. The negative results from Section 5 seem to suggest the necessity of this assumption. Interestingly, and perhaps a little surprisingly, there exist good prequential tests. However, these are tests belonging to a slightly broader category than the ones studied in the previous sections.

Shmaya (2008) shows that there exists such a test T, which passes the truth with probability 1, and cannot be ignorantly passed. This test is, however, not Borel, that is, the sets

$$\{\omega \in \Omega : T(f_1, f_2, ..., \omega) = PASS\}$$

are not Borel. We can still talk about the probability of a theory being accepted or rejected, because these sets have the form of the union of a Borel set B and a subset of a Borel set of measure 0. Therefore, like the

definition of Lebesgue measure, we can extend any probability measure defined on the Borel $\sigma-$algebra by assigning to any such set the measure of set B. Shmaya uses ideas similar to those of Dekel and Feinberg, but in his construction replaces Lusin sets with *universally null sets*. I refer the reader to Shmaya's paper for the details of his construction.

As reported in Olszewski and Sandroni (2009c),[9] Peter Grünwald has shown that there exists a *random test* $T^{\lambda}(Q,\omega)$, where λ is a random variable, whose verdict depends only on forecasts $(f_1, f_2, ...)$ made along the actual sequence of outcomes; this test passes the truth with probability 1, and cannot be ignorantly passed. More precisely, this random test takes as input the observed sequence of outcomes, the sequence of forecasts made along this sequence of outcomes, and the realization of random variable λ; it returns as output a PASS-or-FAIL verdict.

Test T passes the truth with probability 1 for all realizations of λ. And for any random generator of theories ξ, with probability 1 the realization of λ has the property that there exists a topologically large set of data sets $G \subset \Omega$ such that

$$\xi\left(\left\{Q \in \Delta(\Omega) : T^{\lambda}(Q,\omega) = PASS\right\}\right) = 0, \forall\, \omega \in G.$$

Given any value of λ, test T^{λ} is a category test similar to that constructed in the proof of Proposition 5. I again refer the reader to the original paper for the details of this construction.

7 Restricting the class of allowed data-generating processes

One response to the negative results reported in Section 5 is that the set of allowed data-generating processes (or theories) $\Delta(\Omega)$ is too "large" and too "abstract." Stochastic processes studied in many fields of empirical research have much simpler forms, e.g., the outcomes are identically and independently distributed (i.i.d.). In other words, one may argue that the requirement that a test does not reject any true $P \in \Delta(\Omega)$ should be replaced with the requirement that the test does not reject true P's that belong to a smaller class of processes. And indeed, for many classes of processes, it is straightforward to separate informed and ignorant experts. In the deterministic world in which the informed expert knows with certainty the outcome that will occur, one can pass the expert's theory if the outcome predicted by him or her indeed occurs. Similarly, if one believes that the outcomes are i.i.d., a good test asks the expert for the probability distribution over outcomes, and gives the PASS verdict if the observed frequency of outcomes matches the expert's distribution. In fact, good tests exist for many classes of parametric and semi-parametric probabilistic models studied in econometrics.

However, the claim that the negative results are possible only when abstract data-generating processes are allowed does not seem to be fully justified. Olszewski and Sandroni (2008, 2009a,b) show that many

[9] The result has been published in Olszewski and Sandroni, but was suggested to the authors by Peter Grünwald.

of their negative results hold true when one replaces $\Delta(\Omega)$ with the class of exchangeable processes, i.e., processes which can be represented as mixtures of (i.e., probability distributions over) i.i.d. processes.

It is, however, interesting to see for what classes of data-generating processes good tests exist. For example, how large can such a class be? Consider again the one-period setting discussed in Section 3. In this setting, one can only allow for the distributions that assign a probability higher than $1-\varepsilon$ to one of the outcomes. But with many periods, this bound can be much lower than $1-\varepsilon$. Olszewski and Sandroni (2009b) provide a simple test which cannot be ignorantly passed, and which rejects the truth only when the forecasts are often close to fifty-fifty.

More interestingly, Stewart (2011) constructs a prequential test that rejects the truth only when the true probability distribution P has the property that the uniform probability distribution Q, which assigns equal probabilities to 0 and 1 contingent on any previously observed sequence of outcomes, weakly merges with P with positive probability.[10] This set of distributions P is topologically small in the space $\Delta(\Omega)$.

Al Najjar et al. (2010) make the point that the restrictions on the theories that the expert is allowed to submit should not be guided by what seems intuitively abstract, or by what seems large in the set-theoretic sense. Instead, they should be aligned with normative standards, such as those typically expected of scientific theories and statistical models. They formalize this idea by restricting attention to theories that are learnable and predictive.

They assume that any theory is represented as a probability distribution on a set of parameters Θ, with each $\theta \in \Theta$ indexing a stochastic process. These representations are assumed to have the property that, as data accumulate, the expert is eventually able to forecast as if he knew the true parameter θ to any desired degree of precision. In addition, given a parameter θ and an integer t, the outcomes of the next t periods hardly improve predictions of outcomes in the distant future.

Sandroni et al. (1999) show that the class of learnable and predictive theories is "testable." Specifically, there is a finite test T such that: (1) T does not reject any (learnable and predictive) data-generating process; and (2) for any random generator of (learnable and predictive) theories, there is a (learnable and predictive) data-generating process such that the ignorant expert using this random generator of theories is rejected by T with arbitrarily high probability.

Fortnow and Vohra (2009) indicate another kind of restriction: they claim that even if there exist random generators of theories that enable ignorant experts to pass a test, the generators may not be implementable for computational reasons. For example, they construct a finite test T,[11] which can be

[10] Given distributions P and Q, we say that Q weakly merges with P at ω if for every $\delta > 0$, there exists some T such that the probability of ω_t contingent on ω^t is lower than δ for all $t > T$.

The distribution Q is said to weakly merge with P with probability π if

$$P(\omega \in \Omega : Q \text{ weakly merges with } P \text{ at } \omega) = \pi.$$

[11] This test is finite in the stronger sense of Sandroni (2003). Specifically, there exists a number n such that for all P, the

implemented in polynomial time, such that for any $\varepsilon > 0$, for sufficiently large integer m, test T passes the truth with probability $1 - \varepsilon$, and any random generator of theories that can ignorantly pass test T with probability $1 - \varepsilon$ can be used to factor m into prime integers.

The existence of an efficient (i.e., probabilistic polynomial-time) algorithm for factoring composite numbers is generally considered unlikely. For example, many commercially available cryptographic schemes take advantage of this fact. However, the problem of computational restrictions seems to be more complicated than it may appear from Fortnow and Vohra's analysis. In particular, Fortnow and Vohra assume that nature gives the informed expert "on a piece of paper" all the probabilities of future events, which enable her to factorize the required numbers. In practice, the informed expert may know just the method of generating correct forecasts, but may face computational restrictions similar to those faced by the ignorant expert.

Finally, it should be mentioned that Tai-Wei Hu and Eran Shmaya have just announced (in the paper not yet available) that if theories and tests are required to be computable (in the sense that they can be described by Turing machines), then there is a future-independent and prequential test that passes the truth with high probability, and cannot be ignorantly passed.

8 Multiple experts

Some researchers (e.g., Al-Najjar and Weinstein (2008) and Feinberg and Stewart (2008)) argue that the negative results depend crucially on whether a single expert is tested in isolation, or multiple experts are tested at the same time. They point out, however, that some limitations on single-expert testing still have force in the multiple-expert setting.

The idea of comparative testing is very attractive. In practice, true probabilistic models may not exist, but we, nevertheless, use probabilistic models. Some models may be better than others, and with the help of data one may be able to determine which models are better (e.g., by comparing the likelihoods of observed events).

The model in which there is no true data-generating process has not yet been examined. Instead, Al-Najjar and Weinstein and Feinberg and Stewart argue that the possibility of comparative testing reverses some of the negative results.

More precisely, Al-Najjar and Weinstein consider prequential tests whose verdict depends on only a finite number n of forecasts and observed outcomes, and this number n is common for all theories P. They restrict attention to situations in which one of two experts is informed and the other is ignorant, and instead of a PASS-or-FAIL verdict their test indicates the expert which it finds to be informed. (They also allow for the verdict to be inconclusive.) They show that some likelihood tests T^n have the following property:

---

test needs only n outcomes in order to give a verdict.

Proposition 6 *If expert i is informed and truthful, then for every $\varepsilon > 0$, there is an integer K such that for all integers n, data-generating processes P, and random generator of theories ξ^j of expert $j \neq i$, the probability of the event that*

(a) T^n picks expert i, or

(b) the probabilities assigned to outcome 1 (or, equivalently, to outcome 0) by the two forecasts differ by at most ε in all but K periods,

is no lower than $1 - \varepsilon$.[12]

Thus, the only way in which an ignorant expert can pass test T^n is for the expert to provide theories that satisfy condition (b). This seems to be difficult to achieve for all data-generating processes P. However, the test does not guarantee that an ignorant expert will fail it, even in the presence of an informed expert. Moreover, the test is unable to reveal the type of the experts when both of them are ignorant. Indeed, Al-Najjar and Weinstein show that there is no test that cannot be ignorantly passed and that can tell whether there is at least one informed expert. (An analogous result was also obtained by Feinberg and Stewart.)

Feinberg and Stewart define a cross-calibration test, under which m experts are tested simultaneously, and which reduces to the calibration test in the case of a single expert, i.e., when $m = 1$. Intuitively, just as the calibration test checks the empirical frequency of observed outcomes conditional on each forecast, the cross-calibration test checks the empirical frequency of observed outcomes conditional on each profile of forecasts.

They show the following proposition:

Proposition 7 (a) *For every data-generating process P, if an expert predicts according to P, the expert is guaranteed to pass the cross-calibration test with probability 1, no matter what strategies the other experts will use.*

(b) *In the presence of an informed expert, for every random generator of theories ξ of an ignorant expert, the subset of data-generating processes P under which the ignorant expert will pass the cross-calibration test with positive probability is topologically small in the space of data-generating processes P.*[13]

However, this test, like the test of Al-Najjar and Weinstein, may be unable to reveal the type of experts when both of them are ignorant. Feinberg and Stewart modify the cross-calibration test to obtain another test, which they call *strict cross-calibration*, and show that:

Proposition 7 (c) *For any random generator of theories $(\xi^1, ..., \xi^m)$, which are independent random variables, the set of realizations ω on which at least two ignorant experts simultaneously pass the strict cross-calibration test with positive probability is a topologically small set in Ω.*

[12] The probability is measured here according to the product measure $P \times \xi^j$ on the space $\Omega \times \Delta(\Omega)$.

[13] The probability is measured here according to the product measure $P \times \xi$ on the space $\Omega \times \Delta(\Omega)$.

Unlike cross-calibration, however, the strict cross-calibration, rejects the informed expert on some (albeit "small") set of data-generating processes P. Therefore, one may ask whether the reversion of the negative result is caused by allowing for simultaneous testing of multiple experts, or rather by allowing for some possibility of rejecting informed experts. Olszewski and Sandroni (2009a) suggest that this is due to the latter rather than the former reason.[14]

Before their result will be presented, it is important to comment on the condition that random generators of theories are independent random variables. In the interpretation, this assumption means that experts cannot collude. If we allowed for correlated generators, experts could provide identical forecasts, in which case multiple-expert tests could typically be ignorantly passed by virtue of the results for single experts.

I will now describe how Olszewski and Sandroni generalize Proposition 3; I will present their result for two experts, but the generalization to any number of experts is straightforward. They consider only tests that reject theories in finite time. That is, they define *comparative tests* which reject theories in finite time as functions that take as input pairs of theories and yield as output two collections of finite sequences of outcomes (one collection for each expert), and which fail the expert's theory if a sequence from her collection is observed.

A test does not reject the truth if the actual data-generating process is, with high probability, not rejected, no matter what theory is provided by the other expert. A test can be ignorantly passed if both experts can randomly select theories, independent of one another, such that the theory selected by each expert will be rejected only with small probability (no higher than $\varepsilon + \delta$), according to the experts' randomization devices, no matter how the data will unfold. Finally, a test is future-independent if the possibility that any expert's theory is rejected in period m depends only on the data observed up to period m and the predictions made by the theories of both experts up to period m.

Proposition 8 *Every comparative future-independent test which rejects theories in finite time, and which does not reject the truth with probability* $1-\varepsilon$*, can be ignorantly passed with probability* $1-\varepsilon-\delta$*.*

The proof of this result combines the proof of Proposition 7 with a fixed-point argument.

9 Bayesian and decision-theoretic approaches to testing experts

9.1 Bayesian approach

Unlike most of this survey, most of economics assumes that even purely informed agents have some correct prior over future events. One may wonder whether such a prior will help a tester to separate informed from ignorant experts. Of course, a tester equipped with a prior can make forecasts herself without the

[14] Olszewski and Sandroni study only tests which rejects theories in finite time. Therefore, their results have no direct implications regarding calibration tests, which give the verdict at infinity.

help of any expert. So, the tester will find the expert's forecasts valuable only when these forecasts are more precise than her own. And only in this case would testing experts seem to make any sense. However, it seems intuitive that the likelihood-ratio test in which the tester compares her own forecasts with the expert's forecasts should reveal that the informed expert knows more than the tester. And the ignorant expert, who knows no more than the tester, should be exposed to some risk of having low likelihood ratios of her forecasts to the expert's forecasts.

Stewart (2011) formalizes this idea as follows: Suppose the informed expert knows the probability distribution P, and the tester knows only a probability distribution μ over probability distributions. Let $\overline{P}$ be the probability distribution over outcomes induced by μ. That is,

$$\overline{P}(A) = \int_{\Delta(\Omega)} P(A) d\mu$$

for every Borel set $A \subset \Omega$. Given a probability distribution P, let $p(\omega^t)[\omega_t]$ be the probability of outcome ω_t conditional on history ω^t; for our purposes, it will be irrelevant how $f(\omega^t)[\omega_t]$ is defined when the probability of ω^t is zero. The conditional probabilities $\overline{p}(\omega^t)[\omega_t]$ are defined analogously. Let

$$\varepsilon = \int_{\Delta(\Omega)} P\left(\left\{\omega \in \Omega : \sum_{t=1}^{\infty} (p(\omega^t)[\omega_t] - \overline{p}(\omega^t)[\omega_t])^2 \text{ converges}\right\}\right) d\mu.$$

Notice that ε is a number, and is a function of μ. This number may not be small; for example, $\varepsilon = 1$ if distribution μ is degenerated to a distribution P. However, ε is small, or even equal to zero, for many distributions μ (for example for the uniform distribution over i.i.d. processes).[15]

Let $T(P, \omega) = PASS$ if $\overline{p}(\omega^t)[\omega_t] = 0$ for some t. In addition, if $\overline{p}(\omega^t)[\omega_t] > 0$ for all t, then $T(P, \omega) = PASS$ if

$$\liminf_t \frac{p(\omega^1)[\omega_1]}{\overline{p}(\omega^1)[\omega_1]} \cdot \ldots \cdot \frac{p(\omega^t)[\omega_t]}{\overline{p}(\omega^t)[\omega_t]} > 1$$

and

$$\sum_{t=1}^{\infty} (p(\omega^t)[\omega_t] - \overline{p}(\omega^t)[\omega_t])^2 \text{ diverges.}$$

In all other cases $T(P, \omega) = FAIL$.

Stewart shows the following proposition:

Proposition 9 *(i) The test passes the truth with probability* $1 - \varepsilon$*; and (ii) for any random generator of theories* ξ*, the ignorant expert who selects a theory according to* ξ *passes the test with probability* 0*.*

The ignorant expert's probability is evaluated according to the product measure $\mu \times \xi$ on the set of all (P, Q), where P is the data-generating process and Q is the expert's theory.

9.2 Decision-theoretic approach

Echenique and Shmaya (2008), Olszewski and Pęski (2011), and Gradwohl and Salant (2011) make the point that in decision theory, information is only a tool for making better decisions. Of course, there is

[15] This fact is nontrivial, see Stewart for details.

no conflict between this view and the literature on testing experts. Even if the decision problem is not explicitly modelled, one may argue that when we know the expert's type, we are able to make better decisions.

However, the impossibility of separating informed and ignorant experts may depend critically on whether the expert's forecasts play the role of advice for a specific decision problem, or whether one simply wishes to learn the expert's type. Indeed, there are two conceptual differences: (1) a tester (a decision maker) must take some default action even in the absence of any expert, and may not appreciate forecasts that suggest the same (or similar) actions; and (2) if forecasts lead to better decisions, the decision maker may appreciate them, no matter what type of expert provides them. Thus, in an analysis of forecasting in the context of a specific decision problem, it seems legitimate to relax both the requirement that a "good" test should always pass informed experts, and the requirement that it should fail ignorant ones.

The general message of these three papers is that in the decision-theoretic setting, the tester is indeed able to benefit almost fully from the possibility of obtaining the expert's advice if the expert is informed, without losing much if the expert is ignorant. The details of the model, various assumptions, and the statements of results vary across the papers. In addition, the generally positive results coexist with some negative ones (see Olszewski and Pęski (2011)). We will not discuss these papers one by one in the present survey. Nevertheless, in order to give some flavor of this kind of analysis, we will describe one result from Gradwohl and Salant (2011), which may show in the most convincing way the general message of these papers and how they contrast with the literature on testing experts.

Gradwohl and Salant study a model in which the expert observes the realizations of some stochastic process. These realization provide signals about the realizations of the data-generating process. In every period, the decision maker decides whether to bet on the outcome of the data-generating process or stay out. The decision maker has no knowledge of the data-generating process, the expert's process, or any relation between the two.

In every period, the expert provides a prediction that specifies, according to the expert's signal, the maximal expected value of betting in that period and the bet that achieves that expected value. If the decision maker decides not to follow the expert's advice, the period ends and both the decision maker and the expert get the payoff of staying out. Otherwise, the decision maker pays the expert a fixed, exogenous share of the maximal expected value of betting in that period, observes the realization of the data-generating process, and obtains the payoff from betting.

Gradwohl and Salant show that if the expert's process satisfies some condition (e.g., when that process coincides with the data-generating process), then the decision maker has a strategy that approximates the first best, that is, the payoff that the decision maker would obtain if she knew the expert's process herself. The strategy is common for all data-generating processes, and all processes of the expert that satisfy the required condition. The approximation is in terms of the average per-period expected payoff, and for any

given level of approximation, the interaction is assumed to last over a sufficiently long time horizon.

In addition, the decision maker can achieve this goal with a fixed and bounded amount of money, which means that for a sufficiently long time horizon, the amount per period is sufficiently small. Therefore even if the expert is "ignorant," the decision maker will not lose much. Gradwohl and Salant show that this result holds for truthful experts, and that the any strategic behavior of the expert (whose process satisfies the required condition) that improves his own payoff over truthfulness can only increase the payoff of the decision maker.[16]

10 Related topics

10.1 Falsifiability and philosophy of science

The literature on testing experts provides a number of insights, and stimulates a discussion on probabilistic modeling, or more generally, the philosophy of science. Olszewski and Sandroni (2011) take the relation to the philosophy of science more literally, and come up with two conclusions. First, they argue that celebrated falsifiability of Karl Popper has no power to distinguish scientific theories from worthless theories when theories can be produced by strategic experts. Second, they find formal support for the maxim that theories should never be fully accepted, and that they should be rejected when proven inconsistent with the data; this maxim clearly contrasts with an approach that accepts a theory once proven to fit the known data.

More specifically, they define a theory $P \in \Delta(\Omega)$ to be *falsifiable* if, for every history $\omega^k = (\omega_1, ..., \omega_{k-1})$, there is an extension $\omega^n = (\omega_1, ..., \omega_{n-1})$ of ω^k such that

$$P(C(\omega^n)) = 0.$$

Thus, a theory is falsifiable if, after any finite sequence of observed outcomes, there is a finite sequence of outcomes that the theory finds impossible to be observed. This *falsifiability test* rejects nonfalsifiable theories out of hand (i.e., on all data sets), while any falsifiable theory is rejected only at all finite sequences of outcomes which the theory finds impossible. Olszewski and Sandroni show that for every $\varepsilon > 0$, the falsifiability test can be ignorantly passed with probability $1 - \varepsilon$.

Olszewski and Sandroni call tests which rejects theories in finite time, *rejection tests*. Recall that such a test specifies for every theory a collection of finite sequences of outcomes, which sequences (according to the test) contradict the theory; and the test fails the theory if one of these sequences is observed. Similarly, *acceptance tests* specify for every theory a collection of finite sequences of outcomes, which sequences

[16] It is worth pointing out that in Gradwohl and Salant (2011), which is typical for the entire literature, the optimal strategies, or the best responses to the strategies of other agents, may not exist without imposing any conditions on the stochastic processes.

(according to the test) confirm the theory; and the tests pass the theory if one of these sequences is observed.

They show that every acceptance test that, with probability $1-\varepsilon$, does not reject the truth can be ignorantly passed with probability $1-\varepsilon-\delta$, while there exists a rejection test that, with probability $1-\varepsilon$, does not reject the truth but cannot be ignorantly passed with probability higher than ε. Moreover, for any good test T^1, there exists a good rejection test T^2 that is harder than T^1, which means that

$$\{\omega \in \Omega : T^2(P,\omega) = PASS\} \subset \{\omega \in \Omega : T^1(P,\omega) = PASS\}, \forall_{P\in\Delta(\Omega)}.$$

10.2 Gaming performance fees by portfolio managers

One example of experts providing probabilistic forecasts is financial analysts. It is tempting to apply the results that ignorant experts cannot be separated from informed experts to managers of financial institutions, especially since gaming returns by portfolio managers seems to be quite common in practice. For example, it is well-known that treating gains and losses asymmetrically creates incentives for managers for taking excessive risk. Lo (2001) examines a hypothetical situation in which a manager takes short positions in S&P 500 put options that mature in one to three months, and showed that such an approach would have generated very sizable excess returns relative to the market in the 1990s.

However, the difficulty in applying our negative results to financial markets is that managers typically use their forecasts to make investment decisions. Therefore, the analysis of this application seems to be closer to that of Section 9.2 rather than to that of Section 5. Yet, Foster and Young (2010) obtain negative results similar in spirit to those from Section 5.

More specifically, suppose that a benchmark portfolio, such as the S&P 500, generates a sequence of stochastic returns x_t in each of T periods $t = 1, ..., T$, and let r_t denote a risk-free rate in period t. A fund has initial value $s_0 > 0$, which, if passively invested in the benchmark portfolio, would generate the return

$$s_0 \prod_{t=1}^{T} x_t$$

by the end of the period T.

Suppose that a skilled manager can generate a sequence of "higher" returns $m_t x_t$. A compensation contract over T periods is a sequence of functions φ_t, $t = 1, ..., T$, such that φ_t is a function of the realizations of m_s and x_s, $s \leq t$, that is, a function of the return of the benchmark portfolio and the excess return generated by the manager. The functions represent payments to a manager, which are made at the end of each period. The contract can also specify a payment in period 0, and the payments can be negative.

Foster and Young show that there is no compensation contract that separates skilled from ignorant managers. More precisely, for any compensation contract which attracts risk-neutral skilled managers, there is a trading strategy for risk-neutral ignorant managers which yields them a higher expected payoff than the benchmark portfolio.

The argument can be explained as follows: Suppose that $s_0 = 1$, $T = 1$ and the benchmark portfolio yields the risk-free rate r. Since the benchmark return is deterministic, the compensation scheme φ is a function of m. Suppose that the skilled manager is able to generate returns $m^* > 1 + r$. Consider the following strategy of the ignorant manager:

The ignorant manager invests s_0 entirely in the benchmark risk-free asset at the beginning of the period. Just before the end of the period, his capital will be $(1 + r)s_0$. He uses this capital as collateral to buy a lottery in the options market that is realized almost immediately, at the end of the period. The lottery is constructed so as to pay $m^*(1 + r)s_0$ with probability $1/m^*$ and to pay 0 with probability $1 - 1/m^*$.

Since the strategy of an ignorant manager generates only one of the two returns m^* or 0, we need to consider only $\varphi(m^*)$ or $\varphi(0)$, that is, the compensation of the skilled manager and the compensation in the case of bankruptcy. Consider the case in which $\varphi(0) < 0$, that is, the manager is financially penalized in the case of bankruptcy. The case of $\varphi(0) < 0$ is simpler. To deter the ignorant manager, the expected fees earned during the period cannot be positive:

$$\left(\frac{1}{m^*}\right)\varphi(m^*) + \left(1 - \frac{1}{m^*}\right)\varphi(0) \leq 0. \tag{7}$$

In order to make the penalty possible, the amount $(1 + r)^{-1} |\varphi(0)|$ must be held in escrow in a safe asset which earns the risk-free rate, and is paid out to the investors if the fund goes bankrupt.

Now consider the skilled manager who can generate the return m^* with certainty. This manager must also put the amount $(1 + r)^{-1} |\varphi(0)|$ in escrow, because ex ante all managers are treated alike and the investors cannot distinguish between them. However, this involves an opportunity cost for the skilled manager, because by investing $(1 + r)^{-1} |\varphi(0)|$ in her own private fund, she could have generated the return $m^*(1 + r)(1 + r)^{-1} |\varphi(0)| = m^* |\varphi(0)|$. The resulting opportunity cost for the skilled manager is

$$m^* |\varphi(0)| - (1 + r)(1 + r)^{-1} |\varphi(0)| = (m^* - 1) |\varphi(0)| .$$

Therefore, she will not participate if the opportunity cost exceeds the fee, that is, if

$$(m^* - 1) |\varphi(0)| \geq \varphi(m^*). \tag{8}$$

However, inequality (8) is equivalent to inequality (7).

11 References

Al-Najjar, Nabil I. and Jonathan L. Weinstein. 2008. "Comparative Testing of Experts." Econometrica, 76(3): 541-559.

Al-Najjar, Nabil I., Rann Smorodinsky, Alvaro Sandroni, and Jonathan L. Weinstein. 2010. "Testing Theories with Learnable and Predictive Representations." Journal of Economic Theory, 145: 2203-2217.

Blackwell, David. 1956. “An analog of the minimax theorem for vector payoffs.” Pacific Journal of Mathematics, 6: 1–8.

Cesa-Bianchi, Nicolò and Gábor Lugosi. 2006. Prediction, Learning and Games. Cambridge: Cambridge University Press.

Dawid, Alexander P. 1982. “The Well-Calibrated Bayesian.” Journal of the American Statistical Association 77: 605-613.

Dekel, Eddie and Yossi Feinberg. 2006. “Non-Bayesian Testing of a Stochastic Prediction.” Review of Economic Studies, 73(4): 893 - 906.

Echenique, Federico and Eran Shmaya. 2008. “You won’t harm me if you fool me.” mimeo.

Fan, Ky. 1953. “Minimax Theorems.” Proceedings of the National Academy of Science U.S.A., 39: 42-47.

Feinberg, Yossi, and Colin Stewart. 2008. “Testing Multiple Forecasters.” Econometrica, 76(3): 541-582.

Fortnow, Lance J., and Rakesh V. Vohra. 2009. “The Complexity of Forecast Testing.” Econometrica, 77(1): 93-105.

Foster, Dean P. 1999. “A proof of calibration via Blackwell’s approachability theorem." Games and Economic Behavior, 29: 73–78.

Foster Dean P., and Rakesh Vohra. 1997. “Calibrated learning and correlated equilibrium.” Games and Economic Behavior, 21: 40–55.

Foster, Dean P., and Rakesh V. Vohra. 1998. “Asymptotic Calibration.” Biometrika, 85(2): 379-390.

Foster, Dean P., and H. Peyton Young. 2010. "Gaming Performance Fees by Portfolio Managers." Quarterly Journal of Economics 125(4): 1435-1458.

Fudenberg, Drew, and David K. Levine 1999. “An Easier Way to Calibrate.” Games and Economic Behavior, 29: 131-137.

Gilboa, Itzhak and David Schmeidler. 1989. “Maxmin Expected Utility with A Non-Unique Prior.” Journal of Mathematical Economics, 18(2): 141-153.

Gradwohl, Ronen, and Yuval Salant. 2011. “How to Buy Advice.” mimeo

Hart, Sergiu, and Andreu Mas-Colell. 2000. “A Simple Adaptive Procedure Leading to Correlated Equilibrium.” Econometrica, 68: 1127-1150.

Kalai, Ehud, Ehud Lehrer and Rann Smorodinsky. 1999. “Calibrated Forecasting and Merging.” Games and Economic Behavior, 29: 151-169.

Lehrer, Ehud. 2001. “Any Inspection Rule is Manipulable.” Econometrica, 69(5): 1333-1347.

Lo, Andrew W. 2001. “Risk management for hedge funds: introduction and overview.” Financial Analysts’ Journal, 56(6): 16-33.

Mannor, Shie, and Gilles Stoltz. 2010. “A Geometric Proof of Calibration.” Mathematics of Operations Research, 35(4): 721-727.

Murphy, Allan H., and Robert L. Winkler. 1977. "Reliability of Subjective Probability Forecasts of Precipitation and Temperature." Journal of the Royal Statistical Society, Series C, 26: 41-47.

Murphy, Allan H., and Edward S. Epstein. 1967. "Verification of Probabilistic Predictions: A Brief Review." Journal of Applied Meteorology, 6: 748-755.

Olszewski,Wojciech, and Alvaro Sandroni. 2008. "Manipulability of Future- Independent Tests." Econometrica, 76(6): 1437-1466.

Olszewski,Wojciech, and Alvaro Sandroni. 2009a. "Manipulability of Comparative Tests." Proceedings of the National Academy of Sciences U.S.A., 106(13): 5029-5034.

Olszewski, Wojciech and Alvaro Sandroni. 2009b. "Strategic Manipulation of Empirical Tests." Mathematics of Operations Research, 34(1): 57-70.

Olszewski, Wojciech, and Alvaro Sandroni. 2009c. "A Nonmanipulable Test." Annals of Statistics, 37(2): 1013 -1039.

Olszewski, Wojciech, and Alvaro Sandroni. 2011. "Falsifiability." American Economic Review, 101: 788-818.

Olszewski, Wojciech and Marcin Pęski. 2011. The principal-agent approach to testing experts. American Economic Journal: Microeconomics, 3(2): 89-113.

Oxtoby, John C. 1980. Measure and Category. Graduate Texts in Mathematics. Springer Verlag.

Raiffa, Howard. 1961. "Risk, Ambiguity, and the Savage Axioms: Comment." The Quarterly Journal of Economics, 75(4): 690-694.

Sandroni, Alvaro. 2003. "The Reproducible Properties of Correct Forecasts." International Journal of Game Theory, 32(1): 151-159.

Sandroni, Alvaro, Rann Smorodinsky and Rakesh V. Vohra. 1999. "Calibration with Many Checking Rules." Mathematics of Operations Research, 28(1): 141-153.

Shmaya, Eran. 2008. "Many inspections are manipulable." Theoretical Economics, 3(3): 367—382.

Shmaya, Eran and Tai-Wei Hu. 2010. "Describable tests need not be manipulable." mimeo.

Stewart, Colin. 2011. "Nonmanipulable Bayesian Testing.", Journal of Economic Theory, 146: 2029–2041

Vovk, Vladimir, and Glenn Shafer. 2005. "Good Randomized Sequential Probability Forecasting is Always Possible." Journal of the Royal Statistical Society Series B, 67(5): 747 - 763.

Vovk, Vladimir. 2007. "Predictions as statements and decisions." mimeo.