Non-Bayesian Testing of an Expert.[*]

Eddie Dekel[†]and Yossi Feinberg[‡]

February 2005

Abstract

We suggest a test for discovering whether a potential expert is informed of the distribution of a stochastic process. In a non-Bayesian non-parametric setting, the expert is asked to make a prediction which is tested against a single realization of the stochastic process. It is shown that by testing the expert with a "small" set of sequences, the test will assure that any informed expert can pass the test with probability one with respect to the actual distribution. Moreover, for the uninformed non-expert it is very difficult to pass this test, in the sense that for any choice of a "small" set of sequences, only a "small" set of measures will assign positive probability to the given set. This technical result may in fact be of independent interest. Hence "most" measures that a non-expert may predict will lead the non-expert to fail the test. We define small as category 1 sets, described in more detail in the paper. **JEL Classification:** K9

1 Introduction

We consider the problem of testing an expert in an uncertain environment. A decision maker named Alice is trying to decide whether Bob, who is potentially an expert, is informed about

[*] Previous versions of this paper were circulated under the title "A True Expert Knows Which Question Should be Asked." We would like to thank Alvaro Sandroni, Eilon Solan and Bob Wilson for helpful comments. The first author gratefully acknowledges the support of the NSF. The second author gratefully acknowledges the support of the Center for Electronic Business and Commerce.

[†] *Department of Economics, Northwestern University, Evanston, USA and Tel-Aviv University, Tel-Aviv, Israel.* e-mail: dekel@northwestern.edu

[‡] *Stanford Graduate School of Business, Stanford CA 94305-5015.* e-mail: yossi@gsb.stanford.edu

the distribution governing a stochastic process. For example, Alice may want to know if Bob is a qualified economist, where an expert economist would know the distribution of a stochastic process governing key economic indicators. Alice herself is not an expert and we assume that she is completely uninformed and non-Bayesian, in the sense that she does not have a prior distribution over the possible distributions that govern the stochastic process, nor does she have a prior over the probability that Bob is an expert. The question is to what extent, in this non-Bayesian and non-parametric setting, Alice can "test" whether Bob is an expert.

Our motivation is twofold. This issue has been discussed extensively in the "calibration" literature, where it has been argued that Alice cannot successfully test Bob. We discuss the connection to this literature in more detail below. In brief, since we obtain a quite different conclusion than previous work, one motivation for our work is to comment on this literature. The second motivation is because we think the question is of interest. Economists typically adopt a Bayesian and game theoretic perspective where Alice and Bob have beliefs over the world, and Alice constructs incentives to get Bob to tell the truth. We think it is of interest to explore the abstract question regarding to what extent one can evaluate a stochastic theory. Obviously deterministic theories can be tested easily—either what they predict comes to pass or does not. If the "truth" is stochastic, how close can one get to testing a prediction? The most obvious scenario where this question arrises is the testing of non-deterministic scientific theories where an outsider who has no ability to create incentives for the scientist wonders whether the theory can be tested (at least, so to speak, in theory).

The tests we permit are comprised of a question, an answer and a grade. The grade depends on the answer and the information available to the grader. In our case, a question asks Bob for a prediction and the grade that Alice assigns has to be based on the prediction and the realization of the stochastic process. The grade is either pass or fail. Our main result states that there is a "good" test such that if Bob is an expert and knows the distribution governing the process, then Bob passes the test with probability one, and, if Bob has no information about the distribution, then it is "practically" impossible for him to pass the test.

We say a test is good if it satisfies the following two properties

1. A good test must yield a passing grade with probability one whenever Bob is an expert who makes the prediction based on the distribution governing the realization.

2. Bob must fail a good test when he is uninformed. Hence Bob should be asked for a

prediction that would be "practically" impossible to make when he is uninformed.

Property 1 asks that the realization fall into the prediction that Bob makes with probability one (according to the actual distribution used, assuming it is known to Bob). Hence we wish to avoid a type I error of rejecting the hypothesis that Bob is an expert when he actually is. Property 2 requires that the realization does not fit the prediction if Bob does not know the distribution that is actually used. Hence we wish to avoid type II errors of accepting the hypothesis that Bob is an expert when he is not. We cannot satisfy property 2 for all guesses that an uninformed Bob might make—obviously when Bob is lucky enough to guess the true distribution he should not be failed—instead we ask that this hold for a large set of such guesses.

The existing literature provides a collection of troubling negative results for this framework. The notion of calibration was suggested by Dawid (1982) for a non-Bayesian nonparametric setting and extensively studied in Dawid (1985). The idea of being calibrated is that by looking at all the times when Bob predicted that something will happen with probability p, then over time the proportion of times where that something happens should be (close to) p. Based on this notion, Foster and Vohra (1998) were first to show that a test requiring Bob to be calibrated with respect to sequential forecasts, is not a good test. While, as Dawid showed, if Bob is an expert he will pass the test satisfying property 1 above, Foster and Vohra demonstrated that an uninformed expert has a randomized strategy that will assure calibrated forecasts with probability one (with respect to the randomized strategy) no matter what the realization of the process. Hence a test that is based on probabilistic predictions as the realization unveils measured against the empirical distribution conditional on the predictions, fails property 2 and does not distinguish an expert from a fraud. This result has been extensively generalized for elaborate calibration tests that are allowed. These include extensions to tests that add dependencies on histories or conditioning on future properties of the realizations as well as randomized tests. Kalai, Lehrer and Smorodinsky (1999), Fudenberg and Levine (1999), Lehrer (2001) and Sandroni, Smorodinsky and Vohra (2003) have shown that non-experts can satisfy many versions of calibration tests. Recently it was shown by Sandroni (2003) that every test based on sequential probabilistic predictions of next period outcomes is doomed to fail property 2 if it satisfies property 1. These results amount to finding a very large class of tests that are *not* good.[1]

[1] The calibration strategies used by a non-expert in these settings have been extensively used in the game theoretical learning literature; cf. Foster and Vohra (1997) and Fudenberg and Levine (1995). A very general

The existing literature is initially unsettling since asking for calibrated forecasts, or more general sequential predictions, appeals to common practices of statistical inference for stochastic processes. Calibration tests compare the empirical distribution of a realization with the predictions made by Bob. These tests ask for a prediction of a sequential law or phenomena in a realization of the process. This is a natural approach when considering commonly applied stochastic processes, such as i.i.d. or Markov chains. But it turns out not to be well suited for the situation we face here. The initial assumption is that *all* distributions are possible, as far as Alice is concerned. How can Alice come up with a sequential test that will test every possible phenomena along a single realization? Indeed, the existing literature states that this is impossible.

We suggest an alternative approach to testing an expert. Our test has Alice asking Bob for a set of realizations, i.e. Bob is asked to predict an event to which an expert assigns probability one. Alice then considers whether the single realization belongs to the set offered by Bob. She also considers whether the set that Bob is predicting is "small" – whether it would seem impossible for a non-expert to predict this event. If the event is small and the realization agrees with this event then Alice can conclude that Bob is an expert. The crucial feature is that Bob's prediction has to be narrow enough to convince Alice that he is not just guessing, and broad enough to be consistent with his actual view if he is an expert. This stands in contrast with sequential (calibrated) forecasts that give no indication of Bob's expertise.

The idea behind our test is that any unique property of a distribution that demonstrates knowledge of the distribution is known only to whoever knows the distribution. Alice has no way of coming up with such a property without being informed. This is why she needs to ask Bob what the right question is—what unique event corresponds to the distribution known only to an expert. Equivalently, Alice could ask Bob to tell her the distribution and then Alice would construct the test based on that distribution, and our main result states that such a test exists.

The motivation for our construction comes from the Bayesian setting. Let T denote the event "Bob passes the test" and E denote the event "Bob is an expert". We are looking for a test that will, conceptually, lead to $P(E|T) = 1$. While we are in a non-Bayesian setting

class of adaptive procedures and the solutions they generate is studied by Hart and Mas-Colell (2000, 2001) and recently Sandroni and Smorodinsky (2004) defined a solution generated by calibration tests.

we can mimic Bayes' formula:

$$P(E|T) = \frac{P(T \cap E)}{P(T \cap E) + P(T \cap \neg E)}$$

and interpret $P(T \cap \neg E) = 0$ as the "sum" over $\mu \in \Delta(\Omega)$ of $\Pr(\mu)P_\mu(T) = 0$, i.e. the sum over all possible measures of the prior probability of each measure times the probability of passing the test according to the measure. However, since we do not have a prior over $\Delta(\Omega)$ we ask that for "most" distributions μ, the probability of passing a given test when μ is the actual distribution equals zero, i.e. we require $P_\mu(T) = 0$ for all but a small set of measures in $\Delta(\Omega)$. Of course this is only an analogy since we are in a non-Bayesian setting.

We need to define the notion of a "small," or unique, event that corresponds to any distribution known by an expert and formalize what we mean by an uninformed Bob being able to pass the test for only a "small" set of measures. Before we do so, consider the following simple example: Assume that the distribution used assigns probability one to a single realization—a single sequence of values. Suppose that Alice asks Bob for an event and Bob predicts an event consisting of this single sequence. Alice will then observe the realization and see that it matches the exact prediction that Bob made. Given knowledge of the actual distribution Bob had probability one of a realization matching his prediction. If he passes this test then Alice is faced with a successful prediction that singles out one outcome out of a continuum of possible sequences. It definitely seems impossible to make such a prediction and pass the test if you know nothing about the distribution. In fact, the "likelihood" equals one out of a continuum—which should be considered small by any method of measuring the likelihood.

For property 1 to hold, the test has to be passed with high probability when Bob is informed. Hence, in general, we need the test to allow for events that have cardinality of the continuum. We conclude that we cannot define small events as being of small cardinality. On the other hand, we cannot use the notion of the support of a measure (the smallest closed set that has probability one) since a test based on the support can easily fail property 2. For example, the support of any non-degenerate i.i.d. process is the whole space of sequences.[2]

[2] Consider any i.i.d. distribution μ over $\Omega = \{0,1\}^\infty$ where at every stage the probability α of the outcome 1 satisfies $0 < \alpha < 1$. Assume by contradiction that $\bar{\omega} = (\bar{\omega}_1, \bar{\omega}_2, \bar{\omega}_3, ...) \in \Omega \setminus Supp\ \mu \neq \emptyset$. Since $\Omega \setminus Supp\ \mu$ is an open set there exists a finite n such that $T = \{\omega = (\omega_1, \omega_2, \omega_3, ...) \in \Omega | \omega_i = \bar{\omega}_i$ for $i = 1, 2, ..., n\} \subset \Omega \setminus Supp\ \mu$. But since μ is an i.i.d. process we have $\mu(T) = \prod_{i=1}^{n} \left(\bar{\omega}_i^{\alpha}(1-\bar{\omega}_i)^{1-\alpha}\right) > 0$ – a

However, much like our Dirac measure example, these i.i.d. distributions do have unique properties that are assigned probability one. Assume the process is generated by an i.i.d. process where the probability of the outcome 1 equals a at each period, the event "The limit of averages of the realized sequence converges to a" occurs with probability 1 according to the strong law of large numbers. This event seems quite small as "many" distributions one can think of will assign this event zero probability.

Our objective is to generalize and formalize these examples. We need to find a notion of a small event such that every distribution will have a small event that occurs with probability 1. We also need to demonstrate that the notion of smallness we suggest distinguishes the expert from the uninformed. We use the notion of category to capture this property of smallness. A category 1 set is a set which is topologically small, it is defined as a countable union of nowhere dense sets, i.e. sets whose closure has an empty interior. Fortunately, for every probability measure there exists a category 1 set that has probability one with respect to the given distribution. For example, the set of sequences with a limit converging to a is a category 1 set.

We now attempt to justify why a prediction of a category 1 set is indeed indication of expertise. Why do we claim it is "impossible" for a non-expert to make such a prediction and pass the test? Our answer lies in the space of all possible probability measures. We show that for *any* category 1 set of sequences, the set of probability measures that assign positive probability to the set of sequences is itself very small. In other words, a given category 1 set of sequences suggested by an uninformed Bob will cause Bob to fail the test with probability one, with respect to any distribution other than a small set of distributions. Once again we need a notion of smallness, but now for a set of distributions. We show that for any given category 1 set of sequences, the set of probability measures that assign positive probability to the category 1 set, is itself a category 1 set in the space of measures endowed with the weak* topology. For example, the set of measures that assign positive probability to the event "The limit of averages of a sequence converges to 2/3" is a category 1 set of measures. In other words, if an uninformed Bob offers this event, or any other category 1 event, as a prediction, he is likely (in the category sense) to fail.

We conclude that with the notion of category 1 for smallness, we can, using only Bob's prediction and a single realization of the process, determine whether Bob is an expert. Alice should ask Bob to predict a category 1 event, if he does so and the realization is in the

contradiction.

predicted set, then she can regard him as an expert, since suggesting a category 1 set as a prediction without being informed seems practically impossible.

In the following section we prove that for any distribution (over a Polish space) there exists a category 1 set S to which that distribution assigns probability 1 and all but a category 1 set of other Borel probability measures assign to S probability zero. Thus we prove the existence of a good test for any prediction over a sufficiently rich set. In the following section we discuss in more detail the connection to the calibration literature, we (briefly) comment further on why one may view category 1 sets as "small," we discuss the possibility of obtaining "finite approximations" of our result, and we observe that our result actually has implications in non-sequential settings as well.

2 A "Good" (non-Bayesian, non-Parametric) Test of a Stochastic Theory

Let $\Omega = \{0,1\}^\infty$ be the set of all countable sequences of 0's and 1's. Each point $\omega \in \Omega$ is called a realization. We consider the set of possible Borel probability measures over Ω denoted by $\Delta(\Omega)$. For any topological space, denote by $\bar{S}$ and $int(S)$ the closure and the interior of a set S, respectively. A set is called *nowhere dense* when the interior of its closure is empty, i.e. $int\left(\bar{S}\right) = \emptyset$. A *first category set* is a countable union of sets that are nowhere dense. A set is called a *second category set* if it is not a first category set.

We first note the following:

Proposition 1 *For every probability measure $\mu \in \Delta(\Omega)$ there exists a set $S_\mu \subset \Omega$ such that $\mu(S_\mu) = 1$ and S_μ is a first category set.*

Proof. From Theorem 16.5 in Oxtoby (1980) we have that for every non-atomic finite measure μ one can divide Ω into a set of category 1 and a μ-measure zero G_δ set, i.e. a countable intersection of μ-measure zero open sets. This follows from Ω being a metric space with a countable base. If μ has atoms then we can add the (at most) countable set of atoms to the first category set to obtain a first category set with μ-measure one. ■

For a constructive example of such a category 1 set, consider the i.i.d. process in $\Omega = \{0,1\}^\infty$ as described in the introduction where the probability of 1 at each period is given by $0 < \alpha < 1$. Consider the set

$$S = \left\{ \omega = (\omega_1, \omega_2, ...) \in \Omega \mid \lim_{n \to \infty} \sum_{i=1}^{n} \frac{1}{n} \omega_i = \alpha \right\}. \tag{1}$$

Claim 2 *S is a category 1 set that occurs with probability 1.*

Proof. From the strong law of large numbers for Bernoulli trials we have that the probability of the event S according to the i.i.d. process equals 1. We define for every $\varepsilon > 0$ and every n the following set:

$$F_{\varepsilon,n} = \left\{ \omega \in \Omega \mid \text{for all } m \geq n, \ \left| \sum_{i=1}^{m} \frac{1}{m} \omega_i - \alpha \right| < \varepsilon \right\}.$$

If $\omega \notin F_{\varepsilon,n}$ then there exists an $m \geq n$ such that $\left| \sum_{i=1}^{m} \frac{1}{m} \omega_i - \alpha \right| \geq \varepsilon$. Consider the set $G_\omega = \{\bar{\omega} \in \Omega | \bar{\omega}_i = \omega_i \ i = 1, ...m\}$, G_ω is an open set in the product topology and for all $\bar{\omega} \in G_\omega$ we have $\left| \sum_{i=1}^{m} \frac{1}{m} \bar{\omega}_i - \alpha \right| \geq \varepsilon$, hence $G_\omega \subset \Omega \setminus F_{\varepsilon,n}$. Since we have found such an open set for every $\omega \notin F_{\varepsilon,n}$ we conclude that $F_{\varepsilon,n}$ is a closed set. Assume $\varepsilon < \alpha/2$. For every $\omega = (\omega_1, \omega_2, ...) \in F_{\varepsilon,n}$ consider the sequence of points $\left\{\omega^k\right\}_{k=1}^{\infty}$ such that $\omega^k = (\omega_1, ..., \omega_k, 0, 0, ...)$. We have that for all k that $\omega^k \notin F_{\varepsilon,n}$ but $\omega^k \to \omega$, hence $F_{\varepsilon,n}$ is nowhere dense. Since $S \subset \bigcup_{n=1}^{\infty} F_{\varepsilon,n}$ for every $\varepsilon > 0$ we have shown that S is included in a countable union of closed nowhere dense sets and is therefore a category 1 set. ■

To determine just how difficult the test is, i.e. whether it is a good test or not, we ask whether the set of possible distributions under which an uninformed Bob could pass the test with positive probability is small. So we now need a notion of smallness for a set of measures. As mentioned, we again use the notion of first category, this time with respect to the space of probability measures over Ω. We consider the weak* topology on the space of measures $\Delta(\Omega)$.[3]

Our main result, which may be of independent interest, states that the set of measures that assign positive probability to a category 1 subset of Ω is itself a small set of measures.

[3] The weak* topology is the weakest topology possible that assures the continuity of measures as operators, i.e. when integrating over continuous functions of Ω there is convergence of the value of the integration for converging continuous functions.

When the expert offers a category 1 set as his prediction we know that only under a small – category 1 – set of possible distributions will he have any positive probability of passing the test.

Theorem 3 *For every first category set $S \subset \Omega$ the set of measure $M_S = \{\mu \in \Delta(\Omega) | \mu(S) > 0\}$ is a category 1 set in the weak* topology.*

Proof. Let S be a first category set in Ω. We can write $S = \bigcup\limits_{i=1}^{\infty} S_i$ where $int(\bar{S}_i) = \emptyset$. Let $S^n = \bigcup\limits_{i=1}^{n} \bar{S}_i$. Hence $\{S^n\}_{n=1}^{\infty}$ is an increasing sequence of closed sets with empty interior and $S \subset \bigcup\limits_{n=1}^{\infty} S^n$. It suffices to show that the set of measures which assign positive probability to $\bigcup\limits_{n=1}^{\infty} S^n$ is a set of category 1 in the weak* topology over $\Delta(\Omega)$. We define the sets $M_S^n \subset \Delta(\Omega)$ by

$$M_S^n = \{\mu \in \Delta(\Omega) | \mu(S^n) \geq \frac{1}{n}\}. \tag{2}$$

If $\mu \in M_S$ then $\mu(S) > 0$ which implies that there exists $\varepsilon > 0$ such that $\mu(\bigcup\limits_{n=1}^{\infty} S^n) \geq \mu(S) > \varepsilon$ and hence there is an m such that $\mu(S^m) = \mu(\bigcup\limits_{n=1}^{m} S^n) > \varepsilon/2$. Choosing $k > Max\{m, \frac{2}{\varepsilon}\}$ we have that $\mu \in M_S^k$. We conclude that $M_S \subset \bigcup\limits_{n=1}^{\infty} M_S^n$. Hence it suffices to show that each set M_S^n is a category 1 set for their countable union to be a category 1 set. We will actually show that M_S^n is a closed nowhere dense set.[4]

Consider M_S^n for a given n. Let $\mu_i \in M_S^n$ $i = 1, 2, ...$ be a sequence of probability measures converging to a probability measure $\mu \in \Delta(\Omega)$ in the weak* topology. Since S^n is a closed set we have

$$\limsup_{i \to \infty} \mu_i(S^n) \leq \mu(S^n) \tag{3}$$

and in fact convergence is equivalent to (3) holding for all closed sets. In particular we find that $\mu \in M_S^n$. We conclude that M_S^n is closed in the weak* topology. Finally we need

[4] The main property we exploit is that for any strictly positive ε the probability measures that assign at least ε probability to a closed subset of S is nowhere dense. Note that for any non-empty set T (not only category 1 sets) the sets of measures that assign positive probability to T is dense in $\Delta(\Omega)$.

to show that $int(M_S^n) = \emptyset$. Let $\mu \in M_S^n$ we will show that for every open set G that contains μ, we can find a measure $\nu \in G \setminus M_S^n$. From the definition of the weak* topology, for every open set G there exists a continuous function $f : \Omega \to \mathbb{R}$ and an $\varepsilon > 0$ such that the open set

$$G_{\varepsilon,f}(\mu) = \{\lambda \in \Delta(\Omega) | \left| \int\limits_\Omega f(\omega)d\lambda(\omega) - \int\limits_\Omega f(\omega)d\mu(\omega) \right| < \varepsilon\} \tag{4}$$

satisfies $\mu \in G_{\varepsilon,f}(\mu) \subset G$. If f is constant then $G_{\varepsilon,f}(\mu) = \Delta(\Omega)$ and any Dirac measure δ_ω on a point $\omega \in \Omega \setminus S^n$ will satisfy $\delta_\omega \in G \setminus M_S^n$. If f is not constant then there exist points ω_l and ω_h such that $f(\omega_l) \leq \int\limits_\Omega f(\omega)d\mu(\omega) \leq f(\omega_h)$ and $f(\omega_l) < f(\omega_h)$ we define

$$\alpha = \frac{f(\omega_h) - \int\limits_\Omega f(\omega)d\mu(\omega)}{f(\omega_h) - f(\omega_l)}. \tag{5}$$

Since f is continuous and since S^n is nowhere dense there exist points ω_1 and ω_2 such that $\omega_1, \omega_2 \notin S^n$ and

$$Max\{|f(\omega_1) - f(\omega_l)|, |f(\omega_2) - f(\omega_h)|\} < \varepsilon. \tag{6}$$

Consider the measure $\nu = \alpha\delta_{\omega_1} + (1-\alpha)\delta_{\omega_2}$, i.e. the convex combination of the Dirac measures at ω_1 and ω_2. By the choice of these points we have that $\nu(S^n) = 0$ and therefore $\nu \notin M_S^n$. From the definition of ν we have

$$\left| \int\limits_\Omega f(\omega)d\nu(\omega) - \int\limits_\Omega f(\omega)d\mu(\omega) \right| = \left| \alpha f(\omega_1) + (1-\alpha)f(\omega_2) - \int\limits_\Omega f(\omega)d\mu(\omega) \right| \tag{7}$$

which from (6) implies

$$\left| \int\limits_\Omega f(\omega)d\nu(\omega) - \int\limits_\Omega f(\omega)d\mu(\omega) \right| < \left| \alpha f(\omega_l) + (1-\alpha)f(\omega_h) - \int\limits_\Omega f(\omega)d\mu(\omega) \right| + \alpha\varepsilon + (1-\alpha)\varepsilon \tag{8}$$

and together with the definition (5) we have

$$\left|\int_{\Omega} f(\omega)d\nu(\omega) - \int_{\Omega} f(\omega)d\mu(\omega)\right| < \left|\int_{\Omega} f(\omega)d\mu(\omega) - \int_{\Omega} f(\omega)d\mu(\omega)\right| + \varepsilon = \varepsilon \tag{9}$$

and we have shown that $\nu \in G_{e,f}(\mu) \setminus M_S^n \subset G \setminus M_S^n$ and the proof is complete. ■

3 Discussion and extensions

While we have focused on the case where the space of realizations is the space of sequences, we have done so mainly because the existing literature has focused on calibration with respect to a sequence. A closer look at our proofs shows that the result holds for any Polish space Ω. The proof of the proposition requires that the topology on Ω have a countable infinite basis, and the proof of the theorem requires completeness and separability for the weak* topology to be represented by (3). For example, if one has the unit interval as the space of realizations, and Bob could be informed of the distribution over the interval that is used to generate one point from the interval, Alice can use the same method to determine whether Bob is an expert or not by asking him for a category 1 set as a prediction, or asking for the distribution and testing using the appropriate category 1 set.[5]

It is natural to ask to what extent the result holds for finite approximations. Clearly, if Ω is finite the result fails.[6] More generally, as will be clear from the following discussion one cannot bound the size of the set used to test Bob. However, the sequential environment provides a natural approximation question to which the answer is positive. Specifically, the proposition below states that an expert can be tested using a set which occurs with probability *close* to 1 and then Alice would be able to fail a non-expert Bob in *finite* time.

However, Alice cannot determine for sure that Bob is an expert in finite time. For example, if Bob suggests the set of all sequences in $\{0, 1\}^\infty$ that from some point onward have only 0's, then such a countable (and hence category 1) set can never be excluded in finite time. As the result clarifies, the exact time by which Alice can be almost certain whether or not Bob is an expert, may not be known in advance to Alice. It depends on the actual distribution of the sequence in the case that Bob is not an expert.

[5] For instance, to test the uniform distribution on $[0, 1]$ one can test if the realization has (in the limit) half of its digits 1 and half zero.

[6] For instance, if it has n elements and Bob claims the distribution is uniform then should one fail Bob on m outcomes then the chance of rejecting a true expert in that case is m/n and the chance of accepting a false expert is $1 - m/n$.

However, Alice can ask Bob to provide a set of sequences that he is confident will occur with high probability and such that in (unbounded) finite time Bob's expertise will be revealed. Equivalently, Alice can find such a sequence based on any distribution that Bob provides as a stochastic prediction.

Proposition 4 *For every distribution $\mu \in \Delta(\Omega)$ and every $\varepsilon > 0$ there is a closed set with empty interior $S \subset \Omega$ such that $\mu(S) > 1 - \varepsilon$. For every closed set with empty interior $S \subset \Omega$ we have for most distributions λ (other than a category 1 set) that there is a finite N such that with $\lambda-$probability $1 - \varepsilon$ the realization will be outside S within N periods.*

Proof. Given μ we know that there is a first category set that is assigned $\mu-$probability one. Hence there is a finite union of closed sets with empty interior that are assigned probability $1 - \varepsilon$. Denote the finite union by S. Consider any measure λ such that $\lambda(S) = 0$, we know that all but a category 1 collection of measures have this property from our main theorem.

For every $\omega \in \Omega \setminus S$ there exists an open neighborhood that does not intersect S since S is a closed set. In particular, for all $\omega \in \Omega \setminus S$ there exists an $N(\omega)$ such that

$$\{\bar{\omega} | \bar{\omega}_i = \omega_i \text{ for } i = 1, ..., N(\omega)\} \subset \Omega \setminus S.$$

Denote $T_N = \{\omega \in \Omega \setminus S | N(\omega) \leq N\}$. We have that $\bigcup_{N=1}^{\infty} T_N = \Omega \setminus S$. Since $\lambda(\Omega \setminus S) = 1$ there exists an N such that $\lambda(T_N) > 1 - \varepsilon$. We conclude that if Bob reports S as a closed set with empty interior prediction that is based on the actual distribution μ, he will never fail the test with probability of at least $1 - \varepsilon$, but for any prediction of a closed nowhere dense set S, and for most distributions, there is a finite time such that Bob will fail the test within that time with probability of at least $1 - \varepsilon$. ■

At first site there seem to be many differences between our tests and those used in the calibration literature. In the latter Bob is asked in each period to provide a stochastic prediction for the next period outcome and is tested according to that collection of predictions. However, one can consider the following reinterpretation of the results in the literature. Bob can be asked initially at time zero to provide for each period t the stochastic prediction for the following period conditional on every possible history. This is equivalent to asking Bob for the distribution, which as noted above is equivalent to what we do. Allowing Bob to make the predictions sequentially might be intuitive and correspond to some bounded rationality or difficulty in communication or even that Bob is using additional realizations to determine his predictions. But nothing in the calibration literature is based on such features, and in principle the results would apply if Bob were asked to make all his predictions as contingent predictions at date zero.

In the calibration literature Bob is then tested according to his predictions on the realized path. Obviously this is equivalent to saying that for some realizations Bob passes the test and for others Bob does not. So one can think of the calibration literature as asking Bob for a distribution μ, and having a rule determining a set of realizations for each possible distribution, say S_μ^C, under which Bob passes the test. This is equivalent to what we do: the rule is that if Bob predicts μ he is tested according to S_μ. The difference then lies in the way the test is constructed. Calibration tests S_μ^C are quite natural but are too large to be "good". More precisely, the results in this literature state that for the calibration rules considered there is a measure μ such that $S_\mu^C = \Omega$, hence a non-expert suggesting this measure cannot fail.

The "strength" of our test is clearly limited. That is, the category 1 set of distributions that cannot be ruled out may still be big. For instance, if Bob only knows that the distribution on $[0, 1]$ has positive density, then he can claim to "know" that the distribution is uniform and not be failed. This is because the set of measures with positive density is a small set. So the test does evaluate Bob's prediction, but not very precisely. That is, the mapping from distributions over the set of realizations to category 1 subsets of the set of realizations is sufficiently fine for property 2, but it is still coarse in its representation of measures. Bob can choose among many possible category 1 sets, with some strictly more informative than others (by way of inclusion). For example, when the expert knows exactly the sequence that will occur, if the limit of averages converges to a, then Bob could predict the sequence itself, or the set of sequences with an average converging to a, or in fact he can predict the union of any category 1 set with this sequence. Indeed, our framework also does not measure any possible "partial" information that an expert could have by ranking his prediction. Of course, implicitly there is a weak ranking of this form: as in the example above smaller (in terms of set inclusion) category one sets are "better" predictions. More generally, it might be of interest to see if this can be extended to a method of ranking the quality of the prediction. An alternative extension would be to the case where the expert has some information that is not in the form of a distribution. For instance, the expert could know the sequence is determined according to a parametric family of distributions without having a prior belief over this family, from which the distribution is chosen.

Finally, we discuss the extent to which the notion of category 1 for smallness is appropriate. This interpretation is supported by the duality between null-sets and category 1 sets. In Sierpinski (1934) it is shown that under the continuum hypothesis there is a one to one mapping of the interval onto itself such that $f(E)$ is a Lesbeque measure 0 set if and only if E is a first category set.[7] This establishes the following result:

Theorem 5 (Theorem 19.4 **from Oxtoby (1980))** *Consider any proposition involving notions of measure zero, category* 1*, and notions of pure set theory. Under the continuum hypothesis, the proposition holds if and only if the proposition obtained by interchanging the terms "measure zero" and "category* 1*" holds.*

This duality principle suggests that if we cannot use the notion of "measure zero" by using the concept of "category 1" we preserve the same set theoretic deductions. In testing an expert we cannot use the notion of measure zero, precisely because we do not know what the measure is. A prediction needs to be evaluated as "small" independently of any predetermined measure. For the space of measures we can consider various notions for size other than category 1. For instance, one could think that dense sets are large. As pointed out in footnote 4 our category 1 set is dense. In fact, for any fixed $\omega \in \Omega$ the set of probability measures assigning positive probability to ω is dense. Thus saying that a set is large if it is dense seems unreasonable in this context. In economics often *open* and dense sets are considered large, to exclude for instance the set of rationals from being considered large in the unit interval. On the other hand, a set that in nowhere dense is topologically "small". Our category 1 set of measures is an $\aleph_0$ union of nowhere dense sets small sets of probability measures over a space Ω that has the cardinality $2^{\aleph_0}$.[8] Thus, while we cannot claim that category 1 is an unequivocally correct notion of smallness for our purpose, it does provide a "good" (albeit far from ideal) non-Bayesian test for stochastic predictions.

[7] Erdös (1943) generalized this result showing that there is such a mapping f that also satisfies $f = f^{-1}$.

[8] We are grateful to a referee for these comments on the relation between our category 1 sets and dense sets.

References

[1] Dawid, A. P. (1982) "The Well-Calibrated Bayesian." *Journal of the American Statistical Association* **77** (379), 605–613.

[2] Dawid, A. P. (1985) "Calibration-Based Empirical Probability." *The Annals of Statistics* **13** (4), 1251–1274.

[3] Erdös, P. (1943) "Some Remarks on Set Theory." *Ann. of Math.* **44** (2), 643–646.

[4] Foster, D. P. and Vohra, R. V. (1998) "Asymptotic Calibration." *Biometrika* **85** (2), 379–390.

[5] Foster, D. P., and Vohra, R. V. (1997) "Calibrated Learning and Correlated Equilibrium." *Games and Economic Behavior* **21** (1-2), 40–55.

[6] Fudenberg, D., and Levine, D. K. (1995). "Consistency and Cautious Fictitious Play." *Journal of Economic Dynamics and Control* **19**, 1065–1090.

[7] Fudenberg, D., and Levine, D. K. (1999) "Conditional Universal Consistency." *Games and Economic Behavior* **29** (1-2), 104–130.

[8] Hart, S. and Mas-Colell, A. (2000) "A Simple Adaptive Procedure Leading to Correlated Equilibrium." *Econometrica* **68** (5), 1127-1150.

[9] Hart, S., and Mas-Colell, A. (2001) "A general class of adaptive strategies." *Journal of Economic Theory* **98** (1), 26-54.

[10] Kalai, E., Lehrer, E., and Smorodinsky, R. (1999) "Calibrated Forecasting and Merging." *Games and Economic Behavior* **29** (1-2), 151–159.

[11] Lehrer, E. (2001) "Any Inspection Rule is Manipulable." *Econometrica* **69** (5) 1333–1347.

[12] Oxtoby, J. C. (1980) *Measure and category. A survey of the analogies between topological and measure spaces.* Second edition. Springer-Verlag, New York-Berlin.

[13] Sandroni, A. (2003) "The Reproducible Properties of Correct Forecasts." *International Journal of Game Theory* **32** (1), 151–159.

[14] Sandroni, A., and Smorodinsky, R. (2004) “Belief-based Equilibrium.” *Games and Economic Behavior* **47** (1), 157–171.

[15] Sandroni, A., Smorodinsky, R., and Vohra, R. V. (2003) “Calibration with Many Checking Rules.” *Mathematics of Operations Research* **28** (1), 141–153.

[16] Sierpinski, W. (1934) “Sur la dualité entre la première catégorie et la mesure nulle.” *Fund. Math.* **22**, 276–280.