Non-Bayesian Testing of a Stochastic Prediction.*

Eddie Dekel† and Yossi Feinberg‡

December 2005

Abstract

We propose a method to test a prediction of the distribution of a stochastic process. In a non-Bayesian non-parametric setting, a predicted distribution is tested using a realization of the stochastic process. A test associates a set of realizations for each predicted distribution, on which the prediction passes. So that there are no type I errors, a prediction assigns probability 1 to its test set. Nevertheless, these sets are "small", in the sense that "most" distributions assign it probability 0, and hence there are "few" type II errors. It is also shown that there exists such a test that cannot be manipulated, in the sense that an uninformed predictor who is pretending to know the true distribution is guaranteed to fail on an uncountable number of realizations, no matter what randomized prediction he employs. The notion of a small set we use is category I, described in more detail in the paper. **JEL Classification:** K9

1 Introduction

We consider the problem of testing a prediction in an uncertain environment. A decision maker named Alice is trying to decide whether a given distribution determines an observed realization. For example, Alice may want to know if a predicted stochastic process that governs key economic indicators is correct, and furthermore, if the prediction was provided by a self-proclaimed expert, Bob, then Alice may want to evaluate whether Bob is indeed

*Previous versions of this paper were circulated under the title "A True Expert Knows Which Question Should be Asked." We would like to thank John Conlon, Ehud Kalai, Wojciech Olszewski, Gil Reilla, Alvaro Sandroni, Eilon Solan, Bob Wilson, two referees and the editor for helpful comments. The first author gratefully acknowledges the support of the NSF under grant 0111830. 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

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

an expert economist, one who would know the distribution. We assume that Alice herself is non-Bayesian, in the sense that she does not have a prior distribution over the possible distributions that govern the stochastic process. Moreover, if the distribution was provided by Bob she also does not assign a prior probability to his expertise. The question is to what extent, in this non-Bayesian and non-parametric setting, Alice can "test" a given distribution using a single realization, e.g. one sample path of the stochastic process.

Such a test considers the input – a given distribution – and uses the information available to the tester – the single realization – to provide a grade, either pass or fail. Hence, a test maps every possible stochastic prediction to a set of a realizations on which the prediction passes. Ideally, we would like the test to pass a predicted distribution whenever it corresponds to the stochastic process and to fail a prediction that did not yield the realization. Such an ideal test (with no Type I or Type II errors) does not exist as we explain below. But our first positive result, in Proposition 1 of section 2, provides a class of tests that satisfy the following two properties:

1. No type I errors: if the true distribution is the one being predicted, then with probability 1, with respect to this true distribution, the test is passed.

2. Few type II errors: if the wrong distribution is being predicted, then for all but a "small" set of true distributions, the test is failed with probability 1 with respect to the true distribution.

Note that the difference between a test satisfying these properties and an ideal test is the "small" set of true distributions, for a particular prediction, where Type II errors may occur. We formalize this notion of smallness in section 2 below where we present this result.

To obtain such a result a distribution must be tested according to features as unique as possible to that distribution. Thus, on the one hand, we want those features to be realized for sure by that distribution, i.e., the features determine an event that has probability 1 according to this distribution. On the other hand, the event has to be unique enough so as to have 0 probability according to all but a "small" set of alternative distributions.

Our motivation is twofold. 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 it 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).

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.

In addressing the above issue it is important to see whether one can evaluate the validity of a stochastic theory and theorist when no incentives are provided.

Our second motivation is the negative results on non-Bayesian testing of experts in the "calibration" literature. Beginning with a result by Foster and Vohra (1998) this literature considered a variety of classes of tests that seem reasonable, yet all turn out to be susceptible to manipulation in the following sense. For each of these tests it was shown that if Bob can choose a predicted distribution randomly then he can pass the test with probability 1 (with respect to his randomized strategy), regardless of the realization.

How does this strong conclusion relate to our result in section 2? That result is concerned with testing a prediction, not a manipulator. It does immediately imply that our tests cannot be manipulated by a pure strategy. However, as noted, calibration tests are manipulated by mixed strategies—random predictions—and our first result does not apply to those. Nevertheless, it turns out that our approach helps avoid manipulation by mixed strategies as well. Specifically, our second positive result, Proposition 2 of section 3, shows that there exists a subclass of the tests constructed in Proposition 1 that cannot be manipulated even with random predictions. Furthermore, for any pure or randomized prediction, Bob is guaranteed to fail the test on a set of realization which is not "small".

We present, in section 4, finite approximations to both positive results for the case of stochastic processes that are realized over time. Section 5 contains a discussion of related literature, in particular the literature on calibration tests.[1]

2 Good tests

Consider a set of possible realizations, Ω; throughout we focus on the case[2] $\Omega = \{0,1\}^{\aleph_0}$. Let $\Delta(\Omega)$ denote the set of probability measures over Ω, when Ω is endowed with the topology generated by the finite cylinders and with the resulting Borel σ-field, and endow $\Delta(\Omega)$ with the weak* topology. The true distribution determining the realization is denoted by $q \in \Delta(\Omega)$. A test is a function $t : \Delta(\Omega) \longrightarrow 2^{\Omega}$ where a predicted distribution p passes the test t if and only if the realization $\omega \in \Omega$ satisfies $\omega \in t(P)$. Note that we consider testing a prediction p with a single realization. Although ω is a sequence, it is just one observation in a single experiment of a stochastic theory and should not be confused with testing a theory with repeated independent experiments.

As discussed, there are two desiderata for an ideal test.

[1] See Dawid (1982, 1985), Foster and Vohra (1998), Kalai, Lehrer and Smorodinsky (1999), Fudenberg and Levine (1999), Lehrer (2001), Sandroni, Smorodinsky and Vohra (2003) and Sandroni (2003).

[2] As can be seen from the proofs, and as we discuss in section 5 below, the positive results in this and the following section hold for a broader class of environments, including for instance the case where $\Omega = [0,1]$.

- An ideal test makes no type I errors, that is it must pass a correct prediction with probability 1, i.e.,

$$q(t(q)) = 1 \text{ for every } q \in \Delta(\Omega). \tag{1}$$

- An ideal test makes no type II errors, that is, it passes predictions other than the truth with probability 0, i.e.,

$$q\left(t\left(p\right)\right) = 0 \text{ for every } q \neq p, \text{ both in } \Delta\left(\Omega\right). \tag{2}$$

An ideal test must find an identifying feature of the predicted distribution via a set of realizations that will occur if the prediction is correct, and that is sufficiently unique to that distribution, i.e., it corresponds to a sufficiently small set, so that it is not realized with positive probability for any other distribution.

However, an ideal test does not exist. Consider a non-atomic prediction p and any $\omega \in t\left(p\right)$. For $q = \delta_\omega$, the measure assigning probability 1 to ω, we have $q\left(t\left(p\right)\right) = 1$, even though $q \neq p$. One way to alleviate this problem would be to consider an environment with a restricted subset of possible true distributions. For example, assume that the only possible measures are the i.i.d. distributions in $\Delta\left(\Omega\right)$ with parameter α, denoted by $\{p_\alpha : \alpha \in [0,1]\}$. Then setting $t\left(p_a\right)$ equal to the set of realizations (sequences) whose proportion of 1's converges to α has neither type I nor type II errors and hence is ideal (in the sense above) when confined to this class of feasible measures. This leads to a natural question: how far can such results be extended by considering other restricted subsets of measures. This is a different approach from what we explore here, as it presumes some *a priori* knowledge about the feasible set of distributions which is not the goal of this paper.

Instead, we consider all possible distributions and we propose a *good*, albeit not ideal, test where we weaken the second requirement to hold for "most" $q \neq p$. That is, given any prediction p, we ask that all but a "small" set of true distributions assign the test $t\left(p\right)$ zero probability.[3] This amounts to asking, given any prediction p, for a set $t\left(p\right)$ of realizations that would be surprising, i.e., unlikely to occur, given all but a "small" subset of $q \neq p$.

We obtain a good test in two steps. First, we observe that any distribution has a "small" set of realizations that occur with probability 1. Then we show that given any "small" set of realizations, the set of measures that assign it non-zero probability is itself "small". That is, the first step looks for a "small" set in Ω, while the second step says that this yields a "small" set in $\Delta\left(\Omega\right)$.

[3] One might alternatively ask that given any true distribution q, "most" predictions fail the test, so that $q\left(t\left(p\right)\right) = 0$ for all q and most p. However, this allows for a prediction p that passes the test for every q. Thus the test would be useless when faced with such a prediction, which conflicts with our objective. Moreover, such a test would be manipulable using a pure strategy. (Precise definitions of manipulability are in the next section.)

The notion of "small" that we use here is topological: a category I set, which is a countable union of nowhere dense sets (sets whose closure has an empty interior).[4] That is, instead of (2) we ask that

$$q\left(t\left(p\right)\right)=0 \text{ for all } p \text{ and all but a category I set of } q\text{'s.} \tag{3}$$

Proposition 1 *1. There exists a test $t:\Delta\left(\Omega\right)\rightarrow 2^{\Omega}$ s.t. for every $p\in\Delta\left(\Omega\right)$ the set $t\left(p\right)$ is category I and $p\left(t\left(p\right)\right)=1$.*

2. For any category I set $\tau\subset\Omega$, we have $\left\{q:q\left(\tau\right)>0\right\}$ is category I.

The proof is in the appendix. The first part follows closely from Oxtoby (1980). The second part is possibly of independent interest as it confirms that the category I notion of smallness carries over in a natural way from sets to distributions over sets.

This result directly provides a class of good tests. The first part of the proposition guarantees for every p the existence of a category I set that has p-probability 1, and hence has no type I errors, i.e., it satisfies equation (1). The second part of the proposition above implies that any such set has type II errors only on a small set of q's, that is it satisfies equation (3).

For a constructive example of such a category I set, consider the i.i.d. distributions described earlier where the probability of 1 at each period is given by $0<\alpha<1$. As we show in the appendix, the set of sequences with proportion α of 1's is a category I set that occurs with probability 1 (according to the i.i.d. process with parameter α).

Proposition 1 implies that the class of good tests which assign a category I set to each prediction can be used to test a stochastic theory about a stochastic environment. While it cannot definitively say that the theory is correct when it passes the test, it does say that only a "small" set of alternative distributions are plausible. However, when faced with a potential expert whose objective is to pass the test, rather than to convince one of a particular theory, we need to be concerned with strategic randomization of the predictions. That is, while the above results indicates that these tests cannot be manipulated by a pure strategy, they can potentially be manipulated by choosing a mixed strategy over predictions. The next section addresses this issue.

3 Good unmanipulable tests

Following our development of good tests, as described in the preceding section, Olzsewski and Sandroni (2005) argued that there exists a good test that can be manipulated by a random

[4] In section 5 we discuss this notion of size.

prediction, in the sense that with probability 1 (according to the randomized strategy) the test is always passed. This follows a long literature where it is shown that a large class of intuitively appealing calibration tests can be manipulated in this way; we discuss this literature in section 5 below. Thus, while the tests developed in the preceding section are good (in the sense defined) at evaluating deterministic predictions they are potentially manipulable by randomized predictions.

Here we develop the first example of tests with no type I errors that cannot be manipulated. We show that any of our category tests can be modified in a way that rules out manipulation. As they remain category tests they retain both properties 1 and 2 required of a good test.

Formally, a test t can be manipulated if the following condition holds.

$$\text{There exists } \mu \in \Delta(\Delta(\Omega)) \text{ such that, for every } \omega \in \Omega,\ \mu\left(\{p|\omega \in t(p)\}\right) = 1.$$

That is, an uninformed predictor has a randomized strategy for choosing predictions such that for every realization he passes the test with probability 1 (with respect to the randomized strategy). So, a test t cannot be manipulated if for every $\mu \in \Delta(\Delta(\Omega))$ there exists an $\omega \in \Omega$, such that $\mu\left(\{p|\omega \in t(p)\}\right) < 1$. We strengthen this in two directions. We ask for the *guaranteed* failure of a non-expert at some realizations, and that the failure occurs on an uncountable set of points. Formally, a test t *cannot be manipulated on an uncountable set of points* if the following holds:

$$\begin{gathered}\text{For every } \mu \in \Delta(\Delta(\Omega)) \text{ there exists an uncountable set } S \subset \Omega \\ \text{such that for every } \omega \in S \text{ we have } \mu\left(\{p|\omega \in t(p)\}\right) = 0.\end{gathered} \tag{4}$$

In other words, no matter what randomized prediction is made, there is an uncountable set of realizations such that on each of these realizations the randomized prediction will fail with probability 1 (with respect to the randomized prediction).

Proposition 2 *Assume the continuum hypothesis. There exists a good test (i.e., satisfying (1) and (3)) that cannot be manipulated on an uncountable set of points.*

The proof is in the appendix. It proceeds by modifying (any) good test from the preceding section, and therefore derives a class of tests. The proposition establishes the existence of a test which assures that no matter what randomized prediction a predictor employs, he is guaranteed to fail on an uncountable set of points, while the test still passes with probability 1 an expert who predicts the true distribution.

4 Finite approximations

There is naturally an interest in finite approximations when the realization is revealed over time. One would like to be able to determine the outcome of the test without waiting for the whole realization to unfold. A finitely determined test can require that rejection will always occur in finite time, that passing will always occur in finite time, or both. If a test passes (respectively fails) a prediction p in finite time n for a realization $\omega = (\omega_1, \omega_2, ...)$ then the test passes (respectively fails) on every realization with the same first n coordinates as ω. Formally, the test is determined (passes, fails or both) in finite time on ω if for some n it treats all the realizations in the open cylinder $C_n(\omega) = \{\omega' | \omega'_i = \omega_i, i = 1, ..., n\}$ in the same way (i.e., passes p on all or no elements of this cylinder).

This implies that a good test cannot pass distributions in finite time since passing p on an open set such as $C_n(\omega)$ would imply that an open set of predictions in $\Delta(\Omega)$ also pass with positive probability on this set. Thus, if we want to avoid type II errors for all but a "small" set of distributions, we cannot pass the test in finite time.[5] Hence we focus on rejection in finite time. This is consistent with much of the statistics literature where the perspective is that one can reject, but not pass, a hypothesis with a limited data set.

While accepting in finite time does not yield itself to a good test, rejection in finite time does as long as we allow for a small probability of a type I error. To better understand finite-time rejection, consider for example the prediction $p = \delta_\omega$, the Dirac measure at a given realization ω. Consider testing this prediction using $t(p) = \{\omega\}$, i.e., passing the deterministic prediction of ω if and only if ω does occur. This guarantees *rejection* of p in finite time since for every $\omega' \neq \omega$ there exists an n such that $(\omega_1, ..., \omega_n) \neq (\omega'_1, ..., \omega'_n)$. Obviously, this test does not pass p at ω in finite time.

In this spirit, the following result states that we can approximate good tests with ε-good tests that provide rejection in finite time, where an ε-good test allows for type I error with probability of no more than $\varepsilon > 0$, and type II errors continue to occur on at most a small set of distributions.

Proposition 3 *There exists a test t_ε such that:*

1. *For every distribution $q \in \Delta(\Omega)$ the type I error is at most ε, i.e., $q(t_\varepsilon(q)) > 1 - \varepsilon$, and*

2. *For every prediction p, $t_\varepsilon(p)$ is a closed set with empty interior in Ω.*

[5] If one also requires that the test passes and rejects in finite time then—since the test needs to be determined either way for every ω—we have that it must be determined in a uniformly *bounded* time by the compactness of Ω. This amounts to requiring that a limited amount of data provide a clear distinction among all possible predictions in $\Delta(\Omega)$—a hefty burden that Olszweski and Sandroni (2005) show is not feasible.

We have that t_ε rejects p in finite time since for every $\omega' \notin t_\varepsilon(p)$ there is an open neighborhood of ω' not in $t_\varepsilon(p)$ and therefore an n such that $C_n(\omega') \subset \Omega \setminus t_\varepsilon(p)$. The test t_ε satisfies our second desideratum since $t_\varepsilon(p)$ is a closed nowhere dense, hence a category I set.

The test is constructed from the category tests developed above. Specifically, for every p we know that there is a category I set $t(p)$ that is assigned p-probability one. Hence, for the given $\varepsilon > 0$, there is a finite (sub-)union of closed sets with empty interior that are assigned p-probability higher than $1-\varepsilon$. Let $t_\varepsilon(p)$ be this finite union which is therefore itself a closed set with an empty interior. We have thus shown that any good test t constructed above can be restricted to an ε-good test t_ε, that rejects in finite time, with $t_\varepsilon(p) \subseteq t(p)$.

To illustrate such a restriction consider once again a prediction p_α denoting the i.i.d. distribution with a parameter $0 < \alpha < 1$. As noted above, the test $t(p_\alpha) = \{\omega = (\omega_1, \omega_2, ...) | \frac{1}{n}\sum_{i=1}^n \omega_i \underset{n\to\infty}{\longrightarrow} \alpha\}$ is shown in the appendix to be a category I set such that $p_\alpha(t(p_\alpha)) = 1$. For $\varepsilon > 0$, we can obtain an ε-good test for p_α as follows. Fix a sequence $\psi_j \underset{j\longrightarrow\infty}{\searrow} 0$ and consider the test

$$t_\varepsilon(p_\alpha) = \{\omega = (\omega_1, \omega_2, ...) \mid \text{for all } j = 1, 2, ..., \text{ for all } n_j \leq m < n_{j+1}, \left|\sum_{i=1}^m \frac{1}{m}\omega_i - \alpha\right| < \psi_j\} \tag{5}$$

where $(n_j)_{j=1}^\infty$ is some increasing sequence of indices in $\mathbb{N}$. That is, given the prediction p_α, the test t_ε asks whether the average occurrences of 1's is within ψ_1 of α beginning in period n_1, then when period $n_2 > n_1$ is realized it asks whether the average is within a tighter bound, ψ_2, of α, and so on. Clearly, since $\psi_j \searrow 0$, $t_\varepsilon(p_\alpha) \subset t(p_\alpha)$ and in particular $t_\varepsilon(p_\alpha)$ is nowhere dense. The set $t_\varepsilon(p_\alpha)$ is closed since it is the intersection of the cylinders $G_j = \left\{\omega| \text{ for all } m \text{ s.t. } n_j \leq m < n_{j+1}, \ |\sum_{i=1}^m \omega_i/m - \alpha| < \psi_j\right\}$. Lastly, in the appendix we show that for every positive decreasing sequence of ψ_j there is a selection of periods n_j such that $p_\alpha(t_\varepsilon(p_\alpha)) > 1-\varepsilon$.

Since any good test can be restricted to an ε-good test with finitely determined rejection, we can find an ε-good test with finitely determined rejection that cannot be manipulated.

Proposition 4 *Assume the continuum hypothesis. There exists an ε-good test t_ε that is finitely determined and cannot be manipulated. Formally,*

- *For all $q \in \Delta(\Omega)$, $q(t_\varepsilon(q)) = 1-\varepsilon$,*
- *For all $p \in \Delta(\Omega)$, $q(t_\varepsilon(p)) = 0$ for all but a category I set of $q \in \Delta(\Omega)$,*
- *$t_\varepsilon(p)$ is a closed (nowhere dense) set in Ω, hence it has finitely determined rejection,*
- *For every $\mu \in \Delta(\Delta(\Omega))$ there exists an uncountable set S such that for every $\omega \in S$ we have $\mu(\{p|\omega \in t_\varepsilon(p)\}) = 0$.*

The proof of this proposition appears in the Appendix. The proof again generates a class of tests, since any given unmanipulable good test t can be restricted to provide unmanipulable ε-good tests that are rejected in finite time for arbitrary $\varepsilon > 0$.

5 Discussion

5.1 Modeling and interpretation

5.1.1 "Small" sets

Category I and alternative notions We call category I sets "small." For the space of measures one can consider various notions for size other than category I. For instance, one could think that dense sets are large. As pointed out in footnote 11 in the appendix our category I sets in $\Delta(\Omega)$ are 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 just if it is dense seems unreasonable in this context. In economics often open and dense sets are considered large. The opposite is of course a closed nowhere dense set, of which we permit only countable unions. Hence this still seems small in the uncountable space of $\Delta(\Omega)$.[6] We have not explored other notions of small sets in such a space, such as shyness (see Christensen (1972), Hunt, Sauer, and Yorke (1992), and Anderson and Zame (2001)).

The aforementioned notions are all topological and of course our results depend on the choice of topology for $\Delta(\Omega)$ and σ-field for Ω. The product topology and resulting Borel field as the σ-field for Ω seem natural given the intent to consider also finitely determined tests. The weak* topology on $\Delta(\Omega)$ is the weakest topology preserving the continuity of integration over continuous functions. For this reason it is commonly used to study decision making, as then expected payoffs are continuous in expectations (assuming continuous utility functions); but this does not provide any additional support for this topology in our non-Bayesian environment. One feature we find appealing in our approach, is the consistency of our notion of "small" both in Ω and in $\Delta(\Omega)$, but this is appealing for aesthetic, and not any more fundamental, reasons.

An alternative would be to replace the topological approach and use the notion of measure in Ω. This is difficult because we do not know what the true measure is, and would then like to use some "full support" measure, which does not exist on $\Delta(\Omega)$. (However, the notion of shyness provides an alternative construction requiring that a small set must stay measurably small under linear transformation. See Christensen (1972).) In any case, there is a duality between null-sets (Lebesgue measure zero) and category I sets. Sierpinski (1934) showed that under the continuum hypothesis there is a one to one mapping, f, of the interval onto

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

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 I, 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 I" holds.*

This duality principle suggests that if we cannot use the notion of "measure zero" by using the concept of "category I" we preserve the same set theoretic deductions. Thus, while we cannot claim that category I is an unequivocally correct notion of smallness for our purpose, it does seem to provide a "good" (albeit far from ideal) non-Bayesian test for stochastic predictions.

Is the set on which manipulation fails "big"? The preceding discussion naturally leads one to ask, in terms of the non-manipulation result, how large is the set S on which random predictions are guaranteed to fail. We prove that it is uncountable, and in fact our proof shows that it is *not* category I, so it is not small in this sense.

A related question is how large is the set of true distributions q on which a random prediction μ fails. Recall that we showed that for any pure strategy prediction, p, we have $q\left(t\left(p\right)\right)=0$ for all but a category I set of q. How large is the set of measures that assign strictly positive probability to the set on which μ fails, $\left\{q|q\left(\left\{\omega|\mu\left\{p:\omega\in t_{\varepsilon}\left(p\right)=0\right\}\right\}\right)>0\right\}$? While we do not have an answer to this, for any nonatomic q, the set S on which μ is guaranteed to fail is null.[8] While at first sight this might seem quite negative, note that the convex combination of any distribution with atoms with a non-atomic distribution does have atoms. So the non-atomic distributions are small in this sense. Relatedly, this set is known to be shy (see Stinchcombe (2001)).

In any case, a very nice strengthening of our result has been obtained by Olzsewski and Sandroni (2005). Their result does not require the continuum hypothesis and is constructive, and shows that the set on which random predictions fail is the complement of a category I set. Hence, to the extent that our notion of category I sets is compelling as a notion of small, they identify a good test that assures failure of random predictions on a *large* set.

Are the predictions tight? The "strength" of our test is clearly limited. That is, the category I set of distributions that cannot be ruled out may still be larger than we would like. After all, the set of all possible distributions on $\Delta\left(\Omega\right)$ is *very* big, so even sets that

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

are *relatively* small may appear large on their own. Moreover, the tests do not distinguish between coarser and finer predictions. The prediction that the distribution is i.i.d. with probability 0.5 will pass even if the predictor knows a more precise prediction, such as, that the actual sequence alternates each period between 0 and 1. This indicates a natural way to create more refined tests. It is immediate that for any test t that is more restrictive than another test $\bar{t}$, in the sense that for all p we have $t(p) \subset \bar{t}(p)$, if $\bar{t}$ is unmanipulable and satisfies property 2, the same holds for t. This does not enable comparing experts by having them choose the more restrictive successful prediction, since for most measures p there is no minimal category I set that is assigned p-probability 1. However, as is clear from the example of the alternating sequence, this does enable some ranking of predictions.

5.1.2 Generalizing Ω.

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, discussed next, has focused on calibration with respect to a sequence—a realization that is unveiled over time. A closer look at our proofs shows that the result holds for more general spaces Ω.

This fact does make us view our results with some concern. It suggests that using a single data point (out of a sufficiently large set), one can, to some extent, assess the validity of a prediction regarding the distribution that determines that point. We find this a very strong and surprising result, perhaps too strong.

This is why we think there is good reason to explore further the case mentioned earlier, where the set of possible distributions is restricted. As noted, this enables obtaining tests with *no* type I or type II errors, and hence that are non-manipulable. We believe that studying this question will help us understand what it means to test a stochastic prediction.

5.2 The literature on calibration

The existing literature focused on calibration tests (Dawid (1982, 1985)). A typical calibration test asks, as a sequence $\omega = (w_1, w_2, ...) \in \{0,1\}^{\aleph_0}$ is realized, for sequential forecasts for some (finite number of) future periods. For example, such a test considers at each period n the forecasted probability that $w_{n+1} = 1$ conditional on $(w_1, ..., w_n)$ (the coordinates of ω that were revealed up to that period). The test compares these forecasted probabilities to the empirical distribution following such forecasts. For example, one could look at all periods were the conditional probability estimated for the next period was (about) 0.2 and see whether the proportion of 1's in periods immediately following that prediction converges to (about) 0.2.

Although these tests were defined for sequential forecasts along each sequence, they can

be mapped into our framework. The main difference is that the calibration literature considers forecasts for future periods along a realized path, and we ask for *ex ante* predictions. However, if the expert making these forecasts is simply asked to provide the forecast conditional on any finite sequence of future realizations to which he assigns strictly positive probability, then this generates a prediction $p \in \Delta(\Omega)$. Similarly, any $p \in \Delta(\Omega)$ generates along any sequence $(w_1, w_2, ...)$ a sequence of conditional forecasts. Therefore, given a calibration test we can construct a test $t_c : \Delta(\Omega) \to 2^{\Omega}$ by setting $t_c(p)$ equal to the set of states ω such that the sequential conditional probabilities generated from p given ω will pass the given calibration test along ω.[9]

Calibration tests are mostly constructed to have no type I error (see, e.g., Lehrer (2001) who comments about type I errors). However, Foster and Vohra (1998) showed that a calibration test can be manipulated (and as just discussed such manipulation translates into the definition used herein).[10] This surprising result has been extensively generalized to richer classes of calibration tests, including dependencies on histories or conditioning on future properties of the realized sequence as well as randomized tests; see Kalai, Lehrer and Smorodinsky (1999), Fudenberg and Levine (1999), Lehrer (2001), Sandroni, Smorodinsky and Vohra (2003) and Sandroni (2003). This collection of negative results provided increasing classes of tests for which no type I error implies manipulability.

The manipulability of calibration tests is largely due to their continuity as functions of realizations and predictions. To see the intuition, consider the game where nature chooses a realization and the predictor chooses a prediction and either passes or fails based on the test. Since the test (calibration, as well as good tests) requires no type I error, we have that for every mixed strategy of nature p—the true distribution governing the realizations—there is a *pure* strategy by the predictor that assures passing, namely predicting the true distribution p. The MinMax for the predictor must therefore be "pass". If the test employed is sufficiently continuous, the predictor has a *mixed* strategy guaranteeing "pass" no matter what the true distribution chosen by nature is, including Dirac measures, hence for every realization. By contrast, we search for a test that identifies sufficiently unique properties for

[9] Similarly, a randomized sequence of forecasts can be mapped into an ex ante random prediction (using Kolmogorov's theorem), an element in $\Delta(\Delta(\Omega))$, and conversely (see Lehrer (2001, Remark1)). That is, the manipulation result in the literature is often proven and stated for a behavioral strategy sequence of forecasts, and we refer to it in this dissussion as holding for a mixed-strategy prediction.

[10] Due to its focus on manipulation the calibration literature has not studied the extent to which type II errors occur. We have also not worked on that question. However, as an example, consider the simplest calibration test applied to a prediction that the distribution is p_α, an i.i.d. distribution with parameter α. The test checks if the proportion of 1's until period T converges to within $\varepsilon > 0$ of α as $T \longrightarrow \infty$. For any $\varepsilon > 0$ the set of distributions q other than p_α that will satisfy this condition with probability of at least $1 - \varepsilon$ is an open set. However, one could find a metric for which these open sets are small (and get smaller as ε goes to 0). This might suggest an interesting alternative approach to weakening the no type II error property.

each distribution, and we find a test which, even though the MinMax is "pass", cannot be manipulated as it guarantees failure on an uncountable set of pure strategies (realizations). It is the ever varying and unique properties of each prediction that render the MinMax theorem inapplicable.

On the other hand, calibration tests have an advantage: we ask for the prediction up front and select the set of realizations for passing a test accordingly, whereas the calibration literature makes the test depend on the realized sequence. While calibration tests can be casted as our ex-ante tests, we generally cannot go in the other direction. In that sense our test is suitable for testing a random theory of the world, but not for the case where someone "feels in their bones" what the weather will be each subsequent day, i.e. learns future probabilities sequentially as a realization unfolds. The latter person cannot provide an ex-ante prediction, because they don't have a theory of the stochastic nature of the weather, and hence such a person cannot be tested using our approach. By asking for the prediction up front we are able to construct a test suitable for the unique features of the predicted theory.

6 Appendix

Proof of Proposition 1. From Theorem 16.5 in Oxtoby (1980) we have that for every non-atomic measure $p \in \Delta(\Omega)$ one can divide Ω into a set of category I and a p-measure zero G_δ set, i.e. a countable intersection of p-measure zero open sets. This follows from Ω being a metric space with a countable base and since p is a finite measure. We can set $t(p)$ as the category I set for a non-atomic measure p. If p has atoms then we can add the (at most) countable set of atoms to the first category set associated with the non-atomic part of p and obtain a first category set $t(p)$ with p-measure one.

The second part of the proposition deals with a notion of smallness for the space $\Delta(\Omega)$. We consider the weak* topology on the space of probability measures $\Delta(\Omega)$. The weak* topology is the weakest topology 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 continuous functions, i.e. the topology where $p_i \longrightarrow p$ if and only if for every continuous function f we have $\int f dp_i \longrightarrow \int f dp$.

Let $S \subset \Omega$ be any category I set and consider the set of measures $M_S = \{p \in \Delta(\Omega) | p(S) > 0\}$. We need to show that M_S is a category I set in $\Delta(\Omega)$ under the weak* topology. We can write $S = \bigcup_{i=1}^{\infty} S_i$ where $int(\bar{S}_i) = \emptyset$. Let $S^n = \bigcup_{i=1}^{n} \bar{S}_i$. Hence $\{S^n\}_{n=1}^{\infty}$ is an increasing sequence of closed sets with empty interior and $S \subset \bigcup_{n=1}^{\infty} S^n$. It suffices to show that the set

of measures which assign positive probability to $\bigcup_{n=1}^{\infty} S^n$ is a set of category I in $\Delta(\Omega)$. We define the sets $M_S^n \subset \Delta(\Omega)$ by

$$M_S^n = \{p \in \Delta(\Omega) | p(S^n) \geq \frac{1}{n}\}. \tag{6}$$

If $\mu \in M_S$ then $\mu(S) > 0$ which implies that there exists $\varepsilon > 0$ such that $p(\bigcup_{n=1}^{\infty} S^n) \geq p(S) > \varepsilon$ and hence there is an m such that $p(S^m) = p(\bigcup_{n=1}^{m} S^n) > \varepsilon/2$. Choosing $k > Max\{m, \frac{2}{\varepsilon}\}$ we have that $p \in M_S^k$. We conclude that $M_S \subset \bigcup_{n=1}^{\infty} M_S^n$. Hence it suffices to show that each set M_S^n is a category I set for their countable union to be a category I set. We will actually show that M_S^n is a closed nowhere dense set.[11]

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

$$\limsup_{i \to \infty} \; p_i(S^n) \leq p(S^n), \tag{7}$$

and in fact convergence is equivalent to (7) holding for all closed sets. In particular we find that $p \in M_S^n$. We conclude that M_S^n is closed in the weak* topology. Finally we need to show that $int(M_S^n) = \emptyset$. Let $p \in M_S^n$ we will show that for every open set G that contains p we can find a measure in $G \setminus M_S^n$. From the definition of the weak* topology the following sets are a subbase for the topology:

$$G_{\varepsilon,f}(u) = \{v \in \Delta(\Omega) \mid \left| \int_{\Omega} f(\omega) dv(\omega) - \int_{\Omega} f(\omega) du(\omega) \right| < \varepsilon\} \tag{8}$$

for every continuous (and bounded in our case) function f, every $\varepsilon > 0$ and measure u. Hence, finite intersections of such sets are a base and in particular every open set G with $p \in G$ contains a non-empty finite intersection of sets $G_{\varepsilon_1,f_1}(p), G_{\varepsilon_2,f_2}(p), ..., G_{\varepsilon_k,f_k}(p)$. Since the set of probability measures is compact in the weak* topology (cf. Theorem 6.4 in Parthasarathy (1967)) we can apply the Krein–Milman Theorem (cf. Theorem 3.23 in Rudin (1991)). Hence there exists a finitely supported measure $q = \Sigma_{i=1}^{l} \alpha_i \delta_{\omega_i}$ (a convex combination of the extreme points – Dirac measures – of $\Delta(\Omega)$) such that $q \in \bigcap_{j=1}^{k} G_{\varepsilon_2,f_2}(p)$. Since S^n is a closed set

[11] The main property we exploit is that for any strictly positive ε the set of 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 $\Delta(\Omega)$.

with empty interior we can find sequences of realizations $\{\omega_i^r\}_{r=1}^{\infty} \notin S^n$ such that $\omega_i^r \longrightarrow \omega_i$ for all i. In particular, $q^r = \Sigma_{i=1}^{l} \alpha_i \delta_{\omega_i^r}$ is a sequence of measures all assigning probability 0 to the set S^n and converging to $q \in \bigcap_{j=1}^{k} G_{\varepsilon_2,f_2}(p) \subset G$ which implies that there exists an $\bar{r}$ with $q^{\bar{r}} \in G$ and $q^{\bar{r}}(S^n) = 0$, i.e. $q^{\bar{r}} \in G \setminus M_S^n$ and the proof is complete. ■

We now prove the claim, made in presenting the i.i.d. example, that the set of sequences with proportion α of 1's is category I and has probability 1 with respect to the i.i.d. distribution with parameter α. Let

$$S = \left\{ \omega = (\omega_1, \omega_2, ...) \in \Omega \mid \lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{n} \omega_i = \alpha \right\}. \tag{9}$$

Claim 6 *S is a category I 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:

$$F_{\varepsilon,n} = \left\{ \omega \in \Omega \mid \text{for all } m \geq n, \ \left| \sum_{i=1}^{m} \frac{1}{m} w_i - \alpha \right| < \varepsilon \right\}. \tag{10}$$

If $\omega \notin F_{\varepsilon,n}$ then there exists an $m \geq n$ such that $\left| \sum_{i=1}^{m} \frac{1}{m} w_i - \alpha \right| \geq \varepsilon$. Consider the set $G_\omega = \{\bar{\omega} \in \Omega \mid \bar{w}_i = w_i \ i = 1, ...m\}$, G_ω is an open set in the product topology and for all $\bar{\omega} \in G_\omega$ we have $\left| \sum_{i=1}^{m} \frac{1}{m} \bar{w}_i - \alpha \right| \geq \varepsilon$, 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 $\omega = (w_1, w_2, ...) \in F_{\varepsilon,n}$ consider the sequence of points $\left\{\omega^k\right\}_{k=1}^{\infty}$ such that $\omega^k = (w_1, ..., w_k, 0, 0, ...)$. 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 I set. ■

We also made the following claim in the text.

Claim 7 *For every positive decreasing sequence $\psi_j \underset{j\longrightarrow\infty}{\searrow} 0$ and any given $\varepsilon > 0$ we can find a sequence of periods $n_1 < n_2 < ...$ such that the set*

$$G = \{\omega = \ \mid \textit{for all } j = 1, 2, ... \textit{ for all } n_j \leq m < n_{j+1}, \ \left| \sum_{i=1}^{m} \frac{1}{m} w_i - \alpha \right| < \psi_j\}$$

has p_α measure of at least $1-\varepsilon$.

Proof. By the law of large numbers we have that for every $\psi_j > 0$ there exists an n_j such that the sets

$$G_j = \{\omega \mid \text{for all } n_j \leq m \ \left|\sum_{i=1}^{m} \frac{1}{m}\omega_i - \alpha\right| < \psi_j\}$$

satisfy $p_\alpha(G_j) > 1-\frac{\varepsilon}{2^j}$. This follows from observing that for every ω with $\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{n} w_i = \alpha$ there is an $n(\omega)$ with $\left|\sum_{i=1}^{m} \frac{1}{m}\omega_i - \alpha\right| < \psi_j$ for all $m \geq n(\omega)$. Since $G = \bigcap_{j=1}^{\infty} G_j$ we have

$$p_\alpha(G) = p_\alpha(\bigcap_{j=1}^{\infty} G_j) \geq 1 - \sum_{j=1}^{\infty}(1 - p_\alpha(G_j)) > 1 - \sum_{j=1}^{\infty} \frac{\varepsilon}{2^j} = 1-\varepsilon$$

as required. ■

Before we prove Proposition 2 we establish some preliminary results. First we note that any subset of a category I set is also a category I since any subset of a nowhere dense set is nowhere dense. In addition, any countable union of category I sets is a category I set. A set is called a category II set if it is not a category I set. Note, that the complement of any category I set in Ω is a category II set, but the complement of a category II set can be larger than a category I set.

A set L is called *Lusin set* if L is an uncountable set such that every uncountable subset of L is of category II. The existence of a Lusin set in $[0,1]$ was shown by Lusin (1914) under the continuum hypothesis. In fact, every category II set contains a Lusin set (see Proposition 20.1 in Oxtoby (1980)). After proving the no-manipulation result below we show that there exists a Lusin set L in the space $\Omega = \{0,1\}^{\aleph_0}$.

Given a randomized prediction $\mu \in \Delta(\Delta(\Omega))$ we define the measure $\bar{\mu} \in \Delta(\Omega)$ as:

$$\bar{\mu}(E) = \int_{\Delta(\Omega)} p(E) d\mu(p) \tag{11}$$

for every measurable set E. This measure is sometimes refereed to as the "center of gravity" of the measure μ. Note that since Ω is a compact metric space so is $\Delta(\Omega)$ in the weak* topology (cf. Theorem 6.4 in Parthasarathy (1967)). By the definition of the weak* topology we have that for every continuous function $f \in C(\Omega)$ the functional $f(P) = \int_\Omega f(\omega) dp(\omega)$ is a continuous functional on $\Delta(\Omega)$. In particular the continuous functionals on $\Delta(\Omega)$ separate points. From the convexity and compactness of $\Delta(\Omega)$ in the weak* topology we have that the generalized integral $\int_{\Delta(\Omega)} p d\mu(p)$ exists in the sense that for every linear functional Λ on $\Delta(\Omega)$ we have

$$\Lambda(\bar{\mu}) = \int_{\Delta(\Omega)} (\Lambda(p)) d\mu(p) \tag{12}$$

and $\bar{\mu}$ is a probability measure. See Theorems 3.27 and 3.28 in Rudin (1991). Since the

measure $\bar{\mu}$ must satisfy

$$\int_{\Omega} f(\omega)d\bar{\mu} = \int_{\Delta(\Omega)} (\int_{\Omega} f(\omega)dp)d\mu(p) \tag{13}$$

for every continuous function f we have that regularity implies that (11) is well defined.[12] To see this, consider first a closed set E, we have that $\nu(E) = \inf\{\int f d\nu | f \geq \chi_E\}$ where χ_E is the characteristic function of E. In particular, this holds for $\nu = \bar{\mu}$ as well. By regularity $\nu(G) = \sup\{\nu(E)|E \text{ is closed, } E \subset G\}$ for every measurable set G. Hence we have measurability of $p(E)$ for measurable sets E and $\int_{\Delta(\Omega)} p(E)d\mu(p)$ is defined and coincides with $\bar{\mu}$ as required.

Proof of Proposition 2. Fix an arbitrary good test $\bar{t}$ as in Proposition 1 and a Lusin set $L \subset \Omega$. We define the test t as follows:

$$t(p) = (\bar{t}(p) \setminus L) \cup \{\omega \in L | p(\{\omega\}) > 0\}. \tag{14}$$

The test t maps a probability measure p to a set that only contains points from L if these are atoms of the distribution p. We need to show that t as defined in (14) is indeed a good test and that t cannot be manipulated on a set of category II points.

First note that $\bar{t}(p) \setminus L$ is a category I set since it is a subset of the category I set $\bar{t}(p)$. Since p has at most a countable number of atoms the set $\{\omega \in L | p(\{\omega\}) > 0\}$ is countable and a union of a category I set with a countable (hence category I) set is also a category I set. We conclude that $t(p)$ is a category I set.

Since $\bar{t}(p) \cap L$ is a category I set included in the Lusin set L we have that $\bar{t}(p) \cap L$ is a countable set. Hence the set $\bar{t}(p) \setminus L = \bar{t}(p) \setminus (\bar{t}(p) \cap L)$ is measurable since it is the set difference of a measurable set and a countable set. We have

$$p(\bar{t}(p)) = p(\bar{t}(p) \setminus L) + p(\bar{t}(p) \cap L) = p(\bar{t}(p) \setminus L) + \sum_{\{\omega \in \bar{t}(p) \cap L | p(\{\omega\}) > 0\}} p(\{\omega\}) \tag{15}$$

since $\bar{t}(p) \cap L$ is countable. Since $\bar{t}$ is a category test we have $p(\bar{t}(p)) = 1$ and so p has no atoms outside $\bar{t}(p)$ which together with (15) implies

$$p(t(p)) = p(\bar{t}(p) \setminus L) + \sum_{\{\omega \in L | p(\{\omega\}) > 0\}} p(\{\omega\}) = p(\bar{t}(p)) = 1. \tag{16}$$

We have shown that for all p the set $t(p)$ is a category I set and $p(t(p)) = 1$ hence t is a good test.

Consider any randomized prediction $\mu \in \Delta(\Delta(\Omega))$ where we consider $\Delta(\Omega)$ as the measur-

[12] Since Ω is a separable metric space so is $\Delta(\Omega)$ and the Borel probability measures in $\Delta(\Omega)$ and $\Delta(\Delta(\Omega))$ are regular (see Parthasarathy (1967)).

able space generated by the Borel σ-field induced by the weak* topology. Fix $\mu \in \Delta(\Delta(\Omega))$, we now show that there is a category II set of realizations S such that for all $\omega \in S$ we have $\mu(\{p|\omega \in t(p)\}) = 0$. Let $\omega \in L$ be a point in the Lusin set. We first note that the set $\{p|\omega \in t(p)\} \subset \Delta(\Omega)$ is measurable. Since $\omega \in L$ we have that $\omega \in t(p)$ if and only if $p(\{\omega\}) > 0$ by the definition in (14). Hence for every $\omega \in L$

$$\{p|\omega \in t(p)\} = \{p|p(\{\omega\}) > 0\} \tag{17}$$

so $\{p|\omega \in t(p)\}$ is exactly the set of all measures with an atom at ω since ω is in the Lusin set. This set of measures is measurable in $\Delta(\Omega)$ since it is the countable union of the sets $\{p|p(\{\omega\}) \geq 1/n\}$, $n = 1, 2, 3...$ and each set $\{p|p(\{\omega\}) \geq 1/n\}$ is a closed set in the weak* topology since if $p_i \underset{i\to\infty}{\longrightarrow} p$ for $\{p_i\}_{i=1}^{\infty} \subset \{p|p(\{\omega\}) \geq 1/n\}$ then $p_i = \alpha_i\delta_\omega + (1-\alpha_i)q_i$ is a convex combination of a probability measure and the Dirac measure at ω with $\alpha_i \geq 1/n$ for all i. Taking a converging subsequence of both the α_i's and the q_i's (the latter has a converging subsequence by the compactness of $\Delta(\Omega)$ in the weak* topology) we find a limit with an atom of at least size $1/n$ at δ_ω.

The randomized prediction μ will pass the test t when the realization is $\omega \in L$ with positive probability if and only if $\mu(\{p|\omega \in t(p)\}) > 0$. From (17) we have

$$\mu(\{p|\omega \in t(p)\}) = \mu(\{p|p(\{\omega\}) > 0\}) \tag{18}$$

so the randomized prediction will pass the test t at $\omega \in L$ with positive probability only if $\mu(\{p|p(\{\omega\}) > 0\}) > 0$. From the definition of $\bar{\mu}$ in (11) we have that

$$\mu(\{p|p(\{\omega\}) > 0\}) > 0 \text{ implies } \bar{\mu}(\{\omega\}) > 0. \tag{19}$$

The set of realizations ω such that $\bar{\mu}(\omega) > 0$ is countable hence the set $S = L \setminus \{\omega|\bar{\mu}(\omega) > 0\}$ is a category II set and for every $\omega \in S$ we have $\bar{\mu}(\omega) = 0$ which implies that $\mu(\{p|\omega \in t(p)\}) = 0$. We have shown that for every $\mu \in \Delta(\Delta(\Omega))$ there is a category II set satisfying (4) as required. ■

Proof that there exists a Lusin set in Ω. The proof follows by viewing points in $\Omega = \{0,1\}^{\aleph_0}$ as the dyadic (binary) expansion of points in $[0,1]$. We observe that the set of dyadic expansions of the points in a Lusin set $L \subset [0,1]$ must be a Lusin set in $\Omega = \{0,1\}^{\aleph_0}$.

The dyadic expansion is unique for all but a countable set of points in $[0,1]$. Assume by contradiction that the set of dyadic expansions of members of L, which we denote by $\bar{L} \subset \{0,1\}^{\aleph_0}$, is not a Lusin set in Ω. Then we could find an uncountable category I subset of $\bar{L}$ in $2^{\aleph_0}$. It suffices to show that the inverse of the dyadic expansion maps a closed nowhere dense set in Ω to a closed nowhere dense set in $[0,1]$ (hence a countable union of such sets

will be mapped to at most a countable union of such sets). This will show that a category I set is mapped to a category I set and the proof is obtained by contradicting L being a Lusin set since the dyadic expansion and its inverse map uncountable sets to uncountable sets.

Consider a closed set $S \subset \Omega$. Since S is closed under the product topology its map under the inverse of the dyadic expansion is closed; this is because convergence of the dyadic expansion implies convergence in $[0,1]$. We need to show that if S is nowhere dense in Ω its preimage is nowhere dense in the interval. Consider any point in the interval and any open neighborhood of that point. Since the dyadic open intervals generate the same topology generated by open intervals we can find a dyadic interval in the open neighborhood which contains the point. The dyadic interval is open in Ω and hence contains points outside the nowhere dense set S. Hence these points are mapped by the inverse of the dyadic expansion to points in the dyadic interval. We conclude that every point in $[0,1]$ has points from outside the preimage of S in any open neighborhood and the image of S is therefore nowhere dense as required. ■

Proof of Proposition 4. Let t be as in Proposition 2, that is, an unmanipulable good test. Let $t_\varepsilon(p)$ be a closed set with empty interior such that $t_\varepsilon(p) \subset t(P)$ and $p(t_\varepsilon(p)) > 1-\varepsilon$, i.e. the ε-good test as constructed in proposition 3. For every ω and p we have that $\omega \in t_\varepsilon(p)$ implies $\omega \in t(p)$. Hence for every ω we find

$$\{p \in \Delta(\Omega) | \omega \in t_\varepsilon(p)\} \subset \{p \in \Delta(\Omega) | \omega \in t(p)\}. \tag{20}$$

Applying (20) for every $\omega \in S$ where S is the category II set where μ fails as in Proposition 2, we have

$$\begin{aligned} \mu(\{p &\in \Delta(\Omega) | \omega \in t_\varepsilon(p)\}) \leq \\ \mu(\{p &\in \Delta(\Omega) | \omega \in t(p)\}) = 0. \end{aligned} \tag{21}$$

Here the final equality follows from Proposition 2. ■

References

[1] Anderson, R. M. and Zame, W. R. (2001) "Genericity with infinitely many parameters." Adv. Theor. Econ., 1:Art. 1, 64 pp.

[2] Christensen, J. P. R. (1972) "On sets of Haar measure zero in abelian Polish groups", Israel J. Math. **13** , 255-260.

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

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

[5] Dekel, E., and Feinberg, Y. (2004) "Non-Bayesian Testing of an Expert." *Working paper.*

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

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

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

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

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

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

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

[13] Hunt, B., Sauer, T. and Yorke, J. (1992) "Prevalence: a translation-invariant 'almost every' on infinite-dimensional spaces", Bull. Amer. Math. Soc. **27** 217-238; addendum, Bull. Amer. Math. Soc. **28** (1993), 306-307.

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

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

[16] Lusin, N. (1914) "Sur un problèmé de M. Baire ." *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences, Paris* **158** 1258-1261.

[17] Olzsewski, W. and Sandroni, A. (2005) *preprint.*/private communication? I am assuming we will get a title at some point.<——

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

[19] Parthasarathy, K. R. (1967) *Probability measures on metric spaces.* Probability and Mathematical Statistics, No. 3. Academic Press, New York-London.

[20] Rudin, W. (1991) *Functional analysis.* International Series in Pure and Applied Mathematics. McGraw-Hill, New York.

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

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

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

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

[25] Stinchcombe, M. (2001) "The gap between probability and prevalence: Loneliness in vector spaces." Proc. Amer. Math. Soc. **129**, 451-457.