Testable Forecasts

Luciano Pomatto*
Caltech

April 6, 2017

Abstract

Predictions about the future are often evaluated through statistical tests. As shown by recent literature, many known tests are subject to adverse selection problems and are ineffective at discriminating between forecasters who are competent and forecasters who are uninformed but predict strategically.

This paper presents necessary and sufficient conditions under which it *is* possible to discriminate between informed and uninformed forecasters. These conditions have a natural Bayesian interpretation.

It is shown that optimal tests take the form of simple likelihood-ratio tests comparing forecasters' predictions against the predictions of a hypothetical outside observer. The result rests on a novel connection between the problem of testing strategic forecasters and the classical Neyman-Pearson paradigm of hypothesis testing.

*E-mail: luciano@caltech.edu - Division of the Humanities and Social Sciences, Caltech, Pasadena, CA, 91125. I am grateful to Nabil Al-Najjar, Kim Border, Andres Carvajal, Eddie Dekel, Federico Echenique, Johannes Horner, Nicolas Lambert, Wojciech Olszewski, Mallesh Pai, Larry Samuelson, Alvaro Sandroni, Colin Stewart and Max Stinchcombe for their helpful comments, and to the audiences at Yale, ASU, UT Austin, Stanford, UC Davis, and the 5th World Congress of the Game Theory Society. I thank the Cowles Foundation for Research in Economics, where part of this research was completed, for its support and hospitality.

Contents

1 Introduction

Forecasts are often formulated in terms of probability distributions over future events (e.g., "a recession will happen with probability 0.2"). Probabilistic forecasts appear across a wide variety of economic and scientific activities, including the analysis of weather and climate (Gneiting and Raftery, 2005), aggregate output and inflation (Diebold, Tay and Wallis, 1997), epidemics (Alkema, Raftery and Clark, 2007), seismic hazard (Jordan et al., 2011), financial risk (Timmermann, 2000), demographic variables (Raftery et al., 2012) and elections (Tetlock, 2005), among many others.[1]

One practical difficulty with probabilistic forecasts is that they cannot be falsified by casual observation but only through proper statistical tests. From an economic perspective, a key issue is that statistical tests aimed at evaluating forecasters can be subject to adverse selection. Consider, to illustrate, a forecaster who is asked to predict how a stochastic process of interest will evolve over time and will be evaluated by an empirical test comparing his prediction against the realized sequence of outcomes. The forecaster can be either a *true expert,* who knows the actual distribution P generating the data and is willing to report it truthfully, or a *strategic forecaster,* who is uninformed about the process but is interested in passing the test in order to establish a false reputation of competence. Recent literature shows that many tests of interest cannot discriminate between the two.

In their seminal paper, Foster and Vohra (1998) examine the well-known calibration test.[2] They construct a randomized forecasting algorithm that allows an individual to pass the test regardless of how data unfold and without any knowledge of the true data generating process. By employing such an algorithm, an uninformed but strategic forecaster can completely avoid being discredited by the data, thus defeating the purpose of the test.

This surprising phenomenon is not restricted to calibration. Subsequent work emphasizes one critical feature of the calibration test: the fact that it is free of Type-I errors. For any possible true law P generating the data, where P is an arbitrary probability measure defined over sequences of outcomes, an expert who predicts according to P will pass the calibration test with high probability (Dawid (1982)). This remarkable

[1] Corradi and Swanson (2006) and Gneiting and Katzfuss (2014) review the literature on probabilistic forecasts.

[2] Consider a stochastic process that every day can generate two outcomes, say "rain" and "no rain." A forecaster passes the calibration test if for every $p \in [0, 1]$, the empirical frequency of rainy days computed over the days where the forecaster predicted rain with probability p is close to p.

property ensures that the test is unlikely to reject any competent forecaster. However, as shown by Sandroni (2003), once incentives are taken into account, the same property leads to a general impossibility result for testing probabilistic predictions: *any* test that is free of Type-I errors can be passed by a strategic but uninformed forecaster. Sandroni's impossibility result has been further extended in several directions by Shmaya (2008) and Olszewki and Sandroni (2008,2009), among others.

Tests, such as calibration, that are free of Type-I errors do not impose any restriction on the unknown law P generating the process. The starting point of this paper is the observation that such a degree of agnosticism is all but common in economics and statistics. Indeed, most empirical studies posit that data are generated according to a specific model, often fully specified up to a restricted set of parameters. In this paper I take a similar approach to the problem of testing forecasters. In particular, I examine the problem of testing forecasters in the presence of a theory about the data-generating process.

I consider a framework where it is known that the law generating the data belongs to a given set Λ, which represents a theory, or *paradigm*, about the phenomenon under consideration. Accordingly, forecasters are required to provide forecasts belonging to Λ, while predictions incompatible with the paradigm are rejected out of hand. The goal of this paper is to understand under what paradigms it is possible to construct tests that cannot be manipulated.

For the purpose of this paper, paradigms admit multiple interpretations. A paradigm can be seen as a summary of pre-existing knowledge about the problem. It can also represent the set of restrictions imposed on the data-generating process by a scientific theory of interest. It can, alternatively, be interpreted as a normative standard to which the forecasters' predictions must conform in order to qualify as useful.

Classic examples of paradigms include the classes of i.i.d., Markov or stationary distributions. In this paper, in order to make the analysis applicable to a broad class of environments, no a priori restrictions are imposed over paradigms (beyond measurability).

A paradigm Λ is *testable* if it admits a test with the following three features. First, it is unlikely that the test will reject a true expert who knows the correct law in Λ. Therefore, the test must be free of Type-I errors with respect to laws in the paradigm. Second, for any possible strategy that a forecaster might employ to misrepresent his knowledge, there is a law belonging to Λ under which the forecaster will fail the test with high probability. Hence, strategic forecasters are not guaranteed to avoid rejection.

Third, the test returns a decision (acceptance or rejection) in finite time. So, only by adopting testable paradigms is possible to construct tests that do not reject true experts and cannot be manipulated.

A crucial question, then, is which paradigms are testable. As discussed in the next section, the existing literature provides notable instances of testable classes of distributions (see, among others, Al-Najjar, Sandroni, Smorodinsky and Weinstein (2010) and Olszewski and Sandroni (2009)). However, reasonably general conditions under which a paradigm is testable are not known.

The first step of the analysis is a general characterization of testable paradigms. The result characterizes testability as a statement about a hypothetical Bayesian outside observer. Given a paradigm Λ, consider, for the sake of illustration, an analyst, consumer or statistician who is uncertain about the odds of the data generating process, and who is sophisticated enough to express a prior probability μ over the set of possible laws. The prior assigns probability 1 to the paradigm. It is shown that Λ is testable if and only if there exists at least one prior μ such that the observer, by predicting according to the prior, is led to forecasts that are incompatible with any law in the paradigm.

More formally, testability is equivalent to the existence of a prior μ over the paradigm such that the law $\int P d\mu(P)$ obtained by averaging with respect to the prior is sufficiently distant, in the appropriate metric, from every law P in the paradigm. In other terms, the paradigm is testable if and only if simply taking the theory Λ to be true does not exhaust all possible opinions that a Bayesian rational agent can entertain. The result can also be interpreted geometrically as a specific lack of compactness and convexity of the paradigm.

The paper illustrates a number of applications of this characterization. The result is first applied to examine nonmanipulable tests. In Section 4 I show that given *any* testable paradigm, it is without loss of generality to restrict the attention to standard likelihood-ratio tests. Hence, *all* testable paradigms can be subsumed under a single family of statistical tests. Such tests are constructed as follows.

First, the test creates a fictitious Bayesian forecaster. This forecaster serves as a benchmark and is obtained by placing a sufficiently uninformative prior μ over the paradigm. Actual forecasters are then evaluated by comparing their predictions to the forecasts generated by the test. A forecaster passes the test if only if the realized sequence of outcomes was, ex-ante, deemed more likely by the agent than by the fictitious Bayesian forecaster. It is important to note that likelihood-ratio tests form one the most canonical classes of tests and that their properties are well understood.

The results are strengthened to show that likelihood-ratio tests are in fact optimal. To this end, an additional contribution of the paper is to provide a notion of optimality based on a novel ranking among tests. A test T_1 is *less manipulable than* T_2 when, controlling for sample size and for the level of Type-I error with respect to laws in Λ, the probability that a strategic forecaster can guarantee passing the test is lower under T_1 than under T_2. So, less manipulable tests are more effective at screening between informed and uninformed experts. I show that for *any* paradigm there exists a likelihood-ratio test that is less manipulable than any other test. The result provides a foundation for likelihood-ratio tests as a general methodology for testing probabilistic predictions under adverse selection. As explained in the main text, the result is closely related to the celebrated Neyman-Pearson lemma and highlights a novel connection between the problem of testing strategic forecasters and the standard practice of hypothesis testing.

The existing literature cast doubts on the possibility of identifying and testing strategic forecasters. This paper provides foundations for a general and, perhaps intuitive, criterion for identifying competent forecasters: a predictor is recognized as knowledgeable if his or her forecasts results more accurate (in likelihood terms) than the predictions of a Bayesian endowed with an uninformative prior.

The paper presents additional results on the structure of testable paradigms and nonmanipulable tests. In Section 6 it is shown that the results on likelihood-ratio tests continue to hold in the case where forecasters are evaluated based only on their one-step ahead predictions made along the realized path.

In Section 7 I identify two representative families of testable paradigms. Many examples of testable classes of distributions, such as i.i.d. or irreducible Markov, satisfy a version of the strong law of large numbers. In all cases, the paradigm is *identifiable*: the true law of the process can eventually be inferred from the data. Identifiability is a sufficient, but not necessary, condition for testability. It is an important assumption when, in the absence of experts, the goal is to learn the correct law P from the data. Identifiability becomes, however, a strong requirement when the goal is to test a single law P put forward by a forecaster.

I show that any well-behaved testable paradigm Λ contains a non-trivial identifiable subset $\Lambda' \subseteq \Lambda$. Hence, identifiable paradigms represent a *minimal* family of testable paradigms. The characterization introduced in this paper also allows us to characterize *maximal* paradigms, i.e., paradigms that are testable and are not included in any other testable set of distributions. It is shown that maximal paradigms take a particularly

simple form and that any testable paradigm can be enlarged to a maximal one. The result answers an open question posed by Olszewski (2015).

1.1 Related Literature

Following the work of Foster and Vohra (1998), the calibration test has been studied by Foster (1999), Fudenberg and Levine (1999), Kalai, Lehrer and Smorodinsky (1999), Hart and Mas-Colell (2001), Lehrer (2001), Sandroni, Smorodinsky and Vohra (2003), Carvayal (2009), Mannor and Stoltz (2009), and Feinberg and Lambert (2015), among others. The impossibility result of Sandroni (2003) has been extended by Shmaya (2008) and Olszewski and Sandroni (2008).

Dekel and Feinberg (2006) and Olszewski and Sandroni (2009b-2011) provide tests that do not impose any restriction over the data-generating process and are nonmanipulable. These papers consider tests that may not return a decision in any finite time.

Likelihood-ratio tests appear in Al-Najjar and Weinstein (2008) as a method for comparing the predictions of two forecasters under the assumption that at least one of them is informed (see also Feinberg and Stewart (2008) and Olszewski and Sandroni (2009c) for different approaches to testing multiple forecasters). The same type of tests also plays an important role in Stewart (2011). Stewart proposes a Bayesian framework where the tester is endowed with a prior over laws and the forecaster is evaluated according to a likelihood-ratio test against the predictions induced by the prior. In the current paper the tester is not assumed to be Bayesian. Instead, the existence of an appropriate prior which allows to construct a nonmanipulable likelihood-ratio test is shown to be a property that is intrinsic to all testable paradigm.

Al-Najjar, Sandroni, Smorodinsky and Weinstein (2010) consider the set of laws that have a learnable and predictable representation, a class of distributions introduced by Jackson, Kalai and Smorodinsky (1999). They show that the paradigm is testable by constructing a test where experts are asked to announce a deadline after which they must be able to provide sharp predictions about future frequencies of outcomes. The use of a deadline announced by the forecaster is an insight that is also applied in one of the tests in this paper. Olszewski and Sandroni (2009) extend the impossibility result of Sandroni (2003) to any paradigm that is convex and compact and provide examples of testable paradigms.

This paper is also related to the work of Babaioff, Blumrosen, Lambert and Rein-

gold (2010), who consider a principal-agent model where the principal offers a monetary contract with the intent of discriminating between informed and uninformed experts. They show, quite surprisingly, that screening is possible if and only if the true law is restricted to a non-convex set of distributions. There are several important differences between the two approaches. In Blumrosen, Lambert and Reingold (2010) payoffs are a function solely of the monetary transfers (which are allowed to be negative and unbounded). This paper follows the literature on testing strategic experts and emphasizes forecasters' reputational concerns. Hence, transfers are absent and the forecaster expected payoff is the probability of passing the test chosen by the tester. The two papers also arrive at different conclusions. In particular, there exist non-convex paradigms that are not testable, and convex paradigms that are testable.[3]

Other papers have addressed the problem of testing strategic forecasters. Principal-agent models have been studied by Olszewki and Sandroni (2007), Echenique and Shmaya (2007), Gradwohl and Salant (2011), Olszewski and Peski (2011) and Sandroni (2014). Fortnow and Vohra (2009) construct tests that are non-manipulable once computational constraints are taken into account. Hu and Shmaya (2012) show that the paradigm of computable distributions is testable when forecasters are limited to computable strategies. Foster and Vohra (2011) and Olszewski (2015) survey the existing literature.

2 Basic Definitions

In each period an outcome from a finite X is realized. A *path* is an infinite set of outcomes and $\Omega = X^\infty$ denotes the set of all paths. Time is indexed by $n \in \mathbb{N}$, and for each path $\omega = (\omega_1, \omega_2, \ldots)$ the corresponding finite history of length n is denoted by ω^n. That is, ω^n is the set of paths that coincide with ω in the first n periods. In addition, ω^0 denotes the empty history. Throughout the paper $\mathcal{F}_n$ denotes the algebra generated by all histories of length n and $\mathcal{B}$ is the σ-algebra generated by $\bigcup_n \mathcal{F}_n$. The set of paths Ω is endowed with the product topology, which makes $\mathcal{B}$ the corresponding Borel σ-algebra. Given a measurable subset $\Gamma \subseteq \Omega$, denote by $\Delta(\Gamma)$ the set of all Borel probability measures assigning probability 1 to Γ. Elements of $\Delta(\Omega)$ will be interchangeably referred to as *laws* or *distributions*. The space $\Delta(\Omega)$ is endowed with

[3]For instance, the paradigm in Al-Najjar, Sandroni, Smorodinsky and Weinstein (2010) is convex, but testable. A paradigm obtained by removing a (non-degenerate) distribution from the set of all distributions is not convex, but is also not testable.

the weak* topology and the corresponding Borel σ-algebra.[4] The same applies to the space $\Delta(\Delta(\Omega))$ of Borel probability measures over $\Delta(\Omega)$. In what follows, the word "measurable" will always mean "Borel measurable."

2.1 Empirical Tests

A *forecaster* announces a law $P \in \Delta(\Omega)$, under the claim that P describes how the data will evolve. A *tester* is interested in evaluating this claim using a statistical test.

Definition 1 *A* test *is a measurable function* $T : \Omega \times \Delta(\Omega) \to [0, 1]$.

The test compares the realized path ω with the reported law P. The law is accepted if $T(\omega, P) = 1$ and rejected if $T(\omega, P) = 0$. Values strictly between 0 and 1 describe randomized tests in which the forecaster is accepted with probability $T(\omega, P)$. Except for Theorem 3, none of the results are affected by restricting the attention to non-randomized tests. The timing is as follows: (i) At time 0, the tester chooses T; (ii) After having observed T, the forecaster chooses whether or not to participate in the test; (iii) A forecaster who chooses to participate must announce a law P; (iv) Nature generates a path ω; and (v) T reports acceptance or rejection.

Throughout the paper the attention is restricted to tests where a decision is reached in finite time. Following Olszewski (2015), a test T is *finite* if for every law P there exists a time n_P such that $T(\cdot, P)$ is measurable with respect to $\mathcal{F}_{n_P}$. That is, P is accepted or rejected as a function of the first n_P observations. The number n_P is deterministic and known ex-ante. A relevant special case is given by the class of *non-asymptotic* tests, where there exists a single deadline N such that $n_P \leq N$ for every P. While the main focus will be on asymptotic tests, as shown in Section 8.2 many of the results extend immediately to non-asymptotic tests. Notice, in addition, that any finite test can be transformed in a non-asymptotic test by ruling out, as inadmissible, those laws that require more than N data points to be evaluated.

2.2 Strategic Forecasting

The forecaster can be of two possible types. A *true expert* (or informed forecaster) knows the law governing the data generating process and is willing to report it truthfully. A

[4] A sequence (P_n) in $\Delta(\Omega)$ converges to P in the weak* topology if and only if $E_{P_n}[\phi] \to E_P[\phi]$ for every continuous function $\phi : \Omega \to \mathbb{R}$. Given a measure Q, E_Q denotes the expectation operator with respect to Q.

strategic (or uninformed) *forecaster* does not possess any relevant knowledge about the data generating process. His goal is to pass the test in order to establish a false reputation of being competent. Strategic forecasters can produce their predictions using mixed strategies. Formally,

Definition 2 *A* strategy *is a randomization over laws* $\zeta \in \Delta(\Delta(\Omega))$.

The next example shows how a standard likelihood-ratio test can be manipulated by strategic forecasters.

Example 1. *(A manipulable likelihood-ratio test)* The test is specified by a time n and a probability measure $Q \in \Delta(\Omega)$ with full support. The law Q serves as a benchmark against which the forecaster is compared. Given a forecast P and a path ω, the test returns 1 if

$$\frac{P(\omega^n)}{Q(\omega^n)} > 1 \tag{1}$$

and 0 otherwise. Thus, the forecaster passes the test if and only if the realized history is more likely under the forecast P than under the benchmark. The test can be manipulated using the following simple strategy. For each history ω^n of length n, consider the measure $P_{\omega^n} = Q(\cdot|\Omega - \omega^n)$ obtained by conditioning Q on the complement of ω^n. It satisfies

$$P_{\omega^n}(\omega^n) = 0 \text{ and } P_{\omega^n}(\tilde{\omega}^n) > Q(\tilde{\omega}^n) \text{ for any other history } \tilde{\omega}^n \neq \omega^n.$$

Define ζ to be the mixed strategy that randomizes uniformly over all measures of the form P_{ω^n} defined above. Given a history ω^n, a forecaster using strategy ζ will pass the test as long as the law he happens to announce is different from P_{ω^n}. This is an event that under ζ has probability greater or equal to $1 - 2^{-n}$. So, no matter how the data will unfold, even for n relatively small the forecaster can be confident of passing the test with high probability.

The test in Example 1 does not assume any structure on the data-generating process. For example, P is not required to belong to a canonical class of distributions such as Markov or i.i.d. The freedom granted to forecasters of announcing any law allows an uninformed predictor to manipulate the test. We will see that once appropriate restrictions are imposed on the type of laws the forecaster can announce, then even simple likelihood-ratio tests similar to the one considered in example can screen between informed and uninformed forecasters.

2.3 Testable Paradigms

The tester operates under a theory, or *paradigm*, about the data generating process. In this paper a theory is identified with the restrictions it imposes over the law of the process. Formally, a paradigm is a measurable set $\Lambda \subseteq \Delta(\Omega)$, with the interpretation that the data are generated according to some unknown law belonging to Λ. Beyond measurability, no assumptions are imposed on Λ. A paradigm can be defined in many ways. For instance, it can express statistical independence between different variables ("the outcome ω_n realized at time n is independent from the outcome realized at time $n-365$") or it might reflect assumptions about the long run behavior of the process ("the process is ergodic").

Given a paradigm, a basic property a test should satisfy is to not reject informed experts.

Definition 3 *Given a paradigm Λ, a nonrandomized test T does not reject the truth with probability $1-\varepsilon$ if for all $P \in \Lambda$ it satisfies*

$$P(\{\omega : T(\omega, P) = 1\}) \geq 1 - \varepsilon. \tag{2}$$

A test that does not reject the truth is likely to accept an expert who reports the actual law of the data generating process.

Tests such as calibration do not reject the truth with respect to the unrestricted paradigm $\Lambda = \Delta(\Omega)$ (see Dawid, 1982). As shown by Sandroni (2003), any such test can be manipulated. That is, given a finite test T that satisfies property (2) for all $P \in \Delta(\Omega)$, there exists a strategy ζ such that

$$\zeta(\{P : T(\omega, P) = 1\}) \geq 1 - \varepsilon$$

for all paths $\omega \in \Omega$. The strategy allows the forecaster to completely avoid rejection. The result motivates the next definition.

Definition 4 *Given a paradigm Λ, a non-randomized test T is ε-nonmanipulable if for every strategy ζ there is a law $P_\zeta \in \Lambda$ such that*

$$(P_\zeta \otimes \zeta)(\{(\omega, P) : T(\omega, P) = 1\}) \leq \varepsilon.$$

The notation $P_\zeta \otimes \zeta$ stands for the independent product of P_ζ and ζ. A test T is ε-nonmanipulable if for any strategy ζ there is a law P_ζ in the paradigm such that

the forecaster is rejected with probability greater than $1-\varepsilon$. Thus, no strategy can guarantee a strategic forecaster more than an ε probability of passing the test.

As discussed by Olszewski and Sandroni (2008), nonmanipulable tests can dissuade uninformed forecasters from participating in the test. This follows from the idea of modelling a strategic forecaster as an agent facing a decision under uncertainty. Assume that a forecaster who opts not to participate in the test receives a payoff of 0, while a forecaster announcing a law P obtains a payoff that depends on the outcome of the test. If P is accepted then the forecaster is recognized to be knowledgeable and gets a payoff $w > 0$. Conversely, if the law is rejected then the forecaster is discredited and incurs a loss $l < 0$.

Further assume that a strategic and uninformed forecaster makes his decision in accordance with the standard Gilboa-Schmeidler (1989) maxmin criterion where each strategy ζ is evaluated according to the minimum expected payoffs with respect to a set of laws. If such a set equals the paradigm, then for each strategy ζ the expected payoff is

$$\inf_{P\in\Lambda} E_{P\otimes\zeta}\left[wT + l\left(1-T\right)\right] \tag{3}$$

where $E_{P\otimes\zeta}$ denotes the expectation operator with respect to $P_\zeta \otimes \zeta$. If ε is sufficiently small, then the value (3) is negative and so the optimal choice for a strategic forecaster is to not take the test. Therefore, a test that rejects the truth with probability $1-\varepsilon$ and is ε-nonmanipulable can screen between informed and uninformed experts: A true expert finds profitable to participate in the test, while for an uninformed expert it is optimal not to participate.[5]

Definitions 3 and 4 extend immediately to general, randomized, tests. Given a paradigm Λ, a test T *does not reject the truth with probability* $1-\varepsilon$ if for every $P \in \Lambda$ it satisfies $E_P\left[T\left(\cdot, P\right)\right] \geq 1-\varepsilon$, where E_P is the expectation operator associated with P. The test is *ε-nonmanipulable* if for every strategy ζ there is a law $P_\zeta \in \Lambda$ such that $E_{P_\zeta\otimes\zeta}\left[T\right] \leq \varepsilon$.

The next definition, which parallels the definition in Al-Najjar, Sandroni, Smorodinsky and Weinstein (2010), summarizes the properties introduced so far.

Definition 5 *A paradigm Λ is* testable *if for every $\varepsilon > 0$ there exists a finite test T such that:*

[5]Section 8.3 considers a different specification of the set C where uninformed forecasters are less conservative and in (3) the worst case scenario is taken with respect to an arbitrary open ball in the paradigm.

1. *T does not reject the truth with probability $1-\varepsilon$; and*

2. *T is ε-nonmanipulable.*

3 Characterization

This section provides a characterization of testable paradigms. It will be useful in what follows to consider the perspective of a Bayesian outside observer (an analyst, a voter, or a statistician) who is interested in the problem at hand and uncertain about the odds governing the data generating process. The uncertainty perceived by the observer is expressed by a prior probability $\mu \in \Delta(\Gamma)$, where $\Gamma \subseteq \Delta(\Omega)$ is the set of laws the observer believes to be possible. Of main interest is the case where Γ equals (or is close to) the paradigm Λ, so that the observer and the tester have compatible views on the problem. If asked to make forecasts about the future, the observer would predict according to the probability measure defined as

$$Q_\mu(E) = \int_\Gamma P(E)\, d\mu(P) \quad \text{for all } E \in \mathcal{B}. \tag{4}$$

The definition (4) follows the well established approach in Bayesian statistical decision theory of defining a probability measure over the sample space Ω by averaging with respect to the prior.[6]

We can now characterize testable paradigms. Given laws P and Q, let $\|P-Q\| = \sup_{E\in\mathcal{B}} |P(E) - Q(E)|$ denote the (normalized) total-variation distance between the two measures. Given a paradigm Λ, its closure with respect to the weak* topology is denoted by $\overline{\Lambda}$.

Theorem 1 *A paradigm Λ is testable if and only if for every $\varepsilon > 0$ there exists a prior $\mu \in \Delta(\overline{\Lambda})$ such that $\|Q_\mu - P\| \geq 1-\varepsilon$ for all $P \in \Lambda$.*

Consider an outside observer whose prior assigns probability 1 to (the closure of) Λ. The result compares the observer's forecasts with the paradigm. Two polar cases are possible. If $Q_\mu \in \Lambda$, then the observer's prediction cannot be distinguished, ex-ante, from the prediction of an expert who announced Q_μ knowing it was the true law of the process. Theorem 1 is concerned with the opposite case, where the prediction Q_μ is

[6] In the literature, Q_μ is often referred to as a *predictive* probability. Cerreia-Vioglio, Maccheroni, and Marinacci (2013) provide, under appropriate conditions, an axiomatic foundation for the representation (4).

far from *any* possible law P in the paradigm. It shows that a paradigm is testable if and only if there is some observer whose uncertainty about the data generating process leads him to predictions that are incompatible with any law in the paradigm.

Theorem 1 allows us to reformulate the property of testability in terms of more standard concepts such as priors and the total-variation distance between probability measures. We now illustrate this idea in the context of a classic environment where a coin of unknown bias is tossed repeatedly.

Example 2 *(The i.i.d. paradigm)* The set of outcomes is $\{Heads, Tails\}$ and a path is an infinite sequence of tosses. The process follows an i.i.d. distribution P_θ, where $\theta \in [0,1]$ is the probability, in each period, of observing $Heads$ under P_θ. The paradigm is the class $\{P_\theta : \theta \in [0,1]\}$. It is well-known that such a paradigm is testable. To see how this fact relates to Theorem 1, let μ be the prior obtained by taking θ to be uniformly distributed. So, μ satisfies $Q_\mu(E) = \int_0^1 P_\theta(E)\, d\theta$ for every event E. For each value θ, let $E_\theta \subseteq \Omega$ be the set of paths where the limiting empirical frequency of $Heads$ equals θ. Then, by the strong law of large numbers, the event E_θ must have probability 1 under P_θ and probability 0 under $P_{\tilde{\theta}}$ for any $\tilde{\theta}$ different from θ. Because μ has no atoms then, for every fixed θ, the event E_θ must have probability 0 under the law Q_μ. Hence, the total-variation distance between Q_μ and P_θ must equal 1 for every law P_θ in the paradigm. Thus, by Theorem 1, the prior μ guarantees that the paradigm is testable.[7]

As shown in Section 8.1, a similar argument shows that the paradigm of all Markov laws is testable.

Testability of a paradigm is a property which can be formulated geometrically as a lack of compactness and convexity. In order to illustrate this idea we now associate to each paradigm Λ an index $I(\Lambda)$ of its compactness and convexity. The definition is based on notions introduced in the context of general equilibrium theory by Folkmann, Shapley, and Starr (see Starr (1969)).

Given a subset $\Lambda \subseteq \Delta(\Omega)$, let

$$I(\Lambda) = \sup_{Q \in \overline{co}(\Lambda)} \inf_{P \in \Lambda} \|Q - P\|$$

[7]The standard argument to show that the i.i.d. paradigm is testable is to consider a test where the forecaster is asked, at time 0, to predict the frequencies of different outcomes in a sufficiently distant future. By the law of large numbers, the test does not reject the truth. In addition, the test is non-manipulable. While intuitive, this argument, which hinges crucially on the law of large numbers, does not extend to general testable paradigm.

where $\overline{co}(\Lambda)$ is the weak*-closed convex hull of Λ. For each Λ we have $0 \leq I(\Lambda) \leq 1$ by the definition of the total-variation distance. If $I(\Lambda) = 0$, then any law Q in the closed convex hull of the paradigm can be approximated with arbitrary precision by a law P in Λ. In this case, it follows from the results of Olszewski and Sandroni (2009) that any test that does not reject the truth is manipulable. In the opposite case, when $I(\Lambda) = 1$, one can find a law in the closed convex hull of Λ that has distance arbitrarily close to 1 with respect to every law in the paradigm. The next result shows that this is true if and only if the paradigm is testable.

Corollary 1 *A paradigm Λ is testable if and only if it satisfies $I(\Lambda) = 1$.*

4 Nonmanipulable Tests

The characterization provided by Theorem 1 allows to study nonmanipulable tests in a unified way. Given any testable paradigm, it is without loss of generality to restrict the attention to simple likelihood-ratio tests:

Theorem 2 *Let Λ be a testable paradigm. Given $\varepsilon > 0$, let $\mu \in \Delta(\overline{\Lambda})$ be a prior that satisfies $\|Q_\mu - P\| > 1 - \varepsilon$ for all $P \in \Lambda$. There exist positive integers $(n_P)_{P \in \Lambda}$ such that the test defined as*

$$T(\omega, P) = \begin{cases} 1 & \text{if } P \in \Lambda \text{ and } P(\omega^{n_P}) > Q_\mu(\omega^{n_P}) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

does not reject the truth with probability $1 - \varepsilon$ and is ε-nonmanipulable.

Given a law P, the test reaches a decision after n_P observations, where n_P is a constant known in advance. The forecaster passes the test if and only if the history realized at time n_P is strictly more likely under P than under the law Q_μ. Notice that the prior μ is required to be sufficiently uninformative so that the induced law Q_μ is far from every law in the paradigm. As implied by Theorem 1, such a prior exists whenever the paradigm is testable.

The likelihood-ratio test is one of the most applied and well-known statistical tests. It is therefore reassuring that all testable paradigms can be unified under the same family of tests and that such tests are already well understood.

The main idea behind the proof of Theorem 2 is to exploit a formal connection between likelihood-ratio tests and the total-variation distance. To illustrate, let A^P be

the set of paths where a law $P \in \Lambda$ passes the test (5), and consider the difference in probability $P\left(A^P\right) - Q_\mu\left(A^P\right)$. It can be shown that by taking n_P large enough, this difference approximates the distance $\|P - Q_\mu\|$. Hence, the event A^P must have probability higher than $1-\varepsilon$ under P, so the test does not reject the truth with high probability. In addition, A^P must have probability at most ε under Q_μ. Thus, in the hypothetical scenario where the data were generated according to Q_μ, a forecaster would be unlikely to pass the test regardless of what law is announced and, therefore, regardless of whether or not he randomizes his prediction. It follows from this observation and from the fact that Q_μ is a mixture of laws in the paradigm, that against every fixed randomization ζ there must be some law P_ζ in the paradigm against which passing the test is unlikely. That is, the test cannot be manipulated.

5 Optimality of Likelihood Tests: a Neyman-Pearson Lemma

Theorem 2 shows that simple likelihood-ratio tests can screen between informed and uninformed forecasters. However, it leaves open the possibility that such tests are inefficient in the number of observations which they require. A natural question is whether there exist tests that, *for a fixed sample size*, can outperform likelihood-ratio tests in screening between experts and strategic forecasters. We now make this question precise by introducing a novel decision-theoretic ranking among tests.

Definition 6 *Let Λ be a paradigm. Given tests T_1 and T_2, say that T_1* is less manipulable than *T_2 if*

$$\sup_{\zeta\in\Delta(\Delta(\Omega))} \inf_{P\in\Lambda} E_{P\otimes\zeta}\left[T_1\right] \leq \sup_{\zeta\in\Delta(\Delta(\Omega))} \inf_{P\in\Lambda} E_{P\otimes\zeta}\left[T_2\right]. \tag{6}$$

Consider a strategic forecaster who is confronted with a test T and must choose whether or not to undertake the test. As discussed in Section 4, an uninformed forecaster will participate only if the value $\sup_{\zeta\in\Delta(\Lambda)} \inf_{P\in\Lambda} E_{\zeta\otimes P}\left[T\right]$ (which is proportional to the maxmin expected payoff from taking the test) is sufficiently large. The ranking (6) requires that any strategic forecaster who finds optimal not to participate in the test T_2 must also find optimal not to participate in the test T_1. Hence, any uninformed forecaster who is screened out by the test T_2 is also screened out by the test T_1. In other terms, a less manipulable test has a stronger effect in deterring strategic forecasters.

A comparison between tests is more informative when some variables, such as the

required number of observations, are kept fixed. To this end, we call a collection $(n_P)_{P\in\Lambda}$ of positive integers a collection of *testing times* if the map $P \mapsto n_P$ is measurable. A test T is *bounded* by the testing times $(n_P)_{P\in\Lambda}$ if $T(\cdot, P)$ is a function of the first n_P observations. The definition allows for the possibility that different laws may need different sample sizes in order to be properly tested. Finally, given a class $\mathcal{T}$ of tests, we will call a test T *least manipulable in* $\mathcal{T}$ if it belongs to $\mathcal{T}$ and is less manipulable than any other test in the same class.

We can now state the main result of this section.

Theorem 3 *Fix a paradigm Λ, testing times $(n_P)_{P\in\Lambda}$ and a probability $\alpha \in [0,1]$. There exists a prior $\mu^* \in \Delta(\overline{\Lambda})$, thresholds $(\lambda_P)_{P\in\Lambda}$, and a test T^* such that:*

1. *$T^*(\omega, P) = 1$ if $P \in \Lambda$ and $P(\omega^{n_P}) > \lambda_P Q_{\mu^*}(\omega^{n_P})$;*
2. *$T^*(\omega, P) = 0$ if $P \notin \Lambda$ or $P(\omega^{n_P}) < \lambda_P Q_{\mu^*}(\omega^{n_P})$; and*
3. *T^* is least manipulable in the class of tests that are bounded by (n_P) and do not reject the truth with probability α.*

Theorem 3 is a general result illustrating the optimality of likelihood-ratio tests. Given the number of data points n_P that the tester is willing to collect for each forecast P, and given a lower bound α on the probability of accepting a true expert, there exists a likelihood-ratio test that is less manipulable than any other test that satisfies the same constraints.

The result does not demand any assumption on the paradigm, which is not required to be testable. Another difference with the test introduced in Theorem 2 is the use of law-specific thresholds λ_P which allow to adjust the probability of accepting a true expert as a function of the desired level α of Type-I errors.[8]

The result is based on a novel connection between the problem of testing strategic forecasters and the methodology of statistical hypotheses testing. To illustrate this idea, consider the standard problem of testing a null hypothesis P_0 against an alternative hypothesis P_1, where P_0 and P_1 are two given probability measures over paths. To be clear, in such a context a (possibly randomized) *hypothesis test* is a function $\phi : \Omega \to [0,1]$, where $\phi(\omega)$ is the probability of accepting the hypothesis P_0 on path ω.

[8]The proof of Theorem 3 provides a complete description of the test T^* and illustrates how the thresholds and the prior μ^* are computed. In the knife-edge case where $P(\omega^{n_P}) = \lambda_P Q_\mu(\omega^{n_P})$ the test is randomized. The use of randomized tests greatly simplifies the analysis and allows the tester to achieve a probability exactly equal to α of accepting a true expert.

Notice that the test T^* is formally equivalent to a hypothesis test where the law P produced by the expert plays the role of the null hypothesis while the outside observer's prediction Q_{μ^*} plays the role of the alternative. The crucial difference between the test T^* and the standard hypothesis testing framework is that the two "hypotheses" P and Q_{μ^*} are not given exogenously. Rather, P is produced by a possibly strategic forecaster while Q_{μ^*} is chosen by the tester.

The celebrated Neyman and Pearson lemma shows that for any two hypotheses P_0 and P_1, given an upper bound on the probability of Type I error, there exists a likelihood-ratio test between P_0 and P_1 that minimizes the probability of Type II errors. The proof of Theorem 3 applies and extends this fundamental result to the problem of strategic forecasters.

The proof proceeds in two steps. First, the belief μ^* is obtained as the solution of an explicit nonlinear minimization problem over the space of priors. The test T^* is then defined by applying the Neyman-Pearson Lemma to each pair of laws P and Q_{μ^*}. The key step is to show, through a duality argument, that because of the particular choice of μ^*, a test which minimizes the probability of Type-II errors with respect to Q_{μ^*} is also a test that is least manipulable.

6 Tests and Off-Path Predictions

The tests presented so far require the forecaster to provide a completely specified law at time 0. In this section we seek tests that evaluate forecasters based only on their one-step-ahead predictions made along the realized path. This is how most practical tests (including calibration) operate. In addition to added realism, such tests have the advantage of taking into account the critical trade-off experts face between proving their knowledge and revealing it through their forecasts. For instance, a competent weather forecaster who intends to profit from his skills might be willing to provide daily forecasts, but not to reveal the model P behind his predictions. Tests where forecasters are required to reveal the law governing the process effectively ignore this trade-off and, as a result, can dissuade even informed experts from participating in the test.

The main result of this section shows that given any testable paradigm it is without loss of generality to restrict the attention to tests which do not rely on off-path, counterfactual, predictions.

We first introduce the necessary notation. Let $\mathcal{H}^n$ be the set of histories of length

n. Consider a forecaster who is asked, in each period, to provide a prediction about the next outcome. A *forecasting rule* is a function $f : \bigcup_{n=0}^{\infty} \mathcal{H}^n \to \Delta(X)$ which specifies, conditional on every history ω^n, the probability $f(\omega^n)(x)$ of observing outcome x in period $n+1$. So, a forecasting rule describes how the forecaster will predict at each history.

It is convenient to identify forecasting rules with laws. This identification is standard. By Bayes' rule every law P induces a forecasting rule f_P obtained by conditioning P at each history. Given a path ω and a time n, the forecasting rule f_P and the law P are related by the identity

$$P(\omega^n) = \prod_{i=0}^{n-1} f_P\left(\omega^i\right)(\omega_{i+1}). \tag{7}$$

In particular, a truthful expert who knows the law generating the data to be P will predict according to f_P.[9] Conversely, given any forecasting rule f, repeated applications of (9) imply that f defines a law P such that $f = f_P$.

The next definition, which follows Dawid (1982), formalizes the requirement that the test be a function only of the sequential predictions made by the forecaster along the realized path.

Definition 7 *A test T is* prequential *if for every pair of laws P and Q and every path ω, if $f_P(\omega^n) = f_Q(\omega^n)$ for every n, then $T(\omega, P) = T(\omega, Q)$.*

As discussed earlier, prequential tests present several advantages. At the same time, the use of prequential tests brings new difficulties. One complication is that such test limit the information available to the tester.

Another difficulty presented by prequential tests is that they provide no indication, at time 0, of how many observations will be necessary for the test to return a pass or fail decision.

One approach for dealing with this issue is to confine the attention to non-asymptotic tests where a decision is reached after a fixed number of observations that is independent of the forecaster's predictions. This approach is applied in Section 8.2.

The approach we follow in this section is to slightly weaken Definition 7. We consider tests where the forecaster is asked to announce, at time 0, the number observations

[9]The forecasting rule f_P is not uniquely defined when some histories have probability 0 under P. In this case, choose f_P to be any forecasting rule that satisfies (7) for all histories that have positive probability under P.

necessary to test his forecasts. The two approaches lead to qualitatively similar results.

We consider the following test:

Definition 8 *Consider a paradigm Λ and a prior $\mu \in \Delta\left(\overline{\Lambda}\right)$. Fix $\varepsilon > 0$. The forecasts-based likelihood-ratio test $T_{\mu,\varepsilon}$ is defined as*

$$T_{\mu,\varepsilon}(d,\omega,P) = \begin{cases} 1 & \text{if } P(\omega^d) > \frac{1}{\varepsilon} Q_\mu(\omega^d) \\ 0 & \text{otherwise} \end{cases}$$

for all $d \in \mathbb{N}$, $\omega \in \Omega$ and $P \in \Delta(\Omega)$.

The test formalizes the following procedure. Before any data is observed, the forecaster is asked to report a deadline d. Then, in each period from 0 to $d-1$, the agent provides a one-step-ahead forecast (i.e. a probability over the set X of outcomes). Consider a forecaster who adopts a rule f_P. At time d, given a path ω, the forecaster is accepted by the test if and only if the ratio

$$\frac{P(\omega^d)}{Q_\mu(\omega^d)} = \prod_{i=0}^{d-1} \frac{f_P(\omega^i)(\omega_{i+1})}{f_{Q_\mu}(\omega^i)(\omega_{i+1})}$$

is above $1/\varepsilon$, where the equality follows directly from (7). For small ε, in order for the forecaster to pass the test, his predictions must accumulate a consistent advantage in likelihood terms over the predictions of a Bayesian endowed with a prior μ. Notice that once a deadline is fixed, the test $T_{\mu,\varepsilon}$ becomes a standard prequential test.[10]

6.1 Main Result

We now study the properties of the forecasts-based likelihood-ratio test. In the context of this test, a *strategy* is a joint randomization $\zeta \in \Delta(\mathbb{N} \times \Delta(\Omega))$ over deadlines and laws.

Theorem 4 *Let Λ be a testable paradigm. Given $\varepsilon > 0$, let $\mu \in \Delta\left(\overline{\Lambda}\right)$ be a prior that satisfies $\|Q_\mu - P\| > 1 - \varepsilon^2/(1+\varepsilon)$ for every $P \in \Lambda$. Then, the test $T_{\mu,\varepsilon}$ satisfies:*

1. *For every P in Λ there is a deadline d_P such that for every $d \geq d_P$*

$$P(\{\omega : T_{\mu,\varepsilon}(d,\omega,P) = 1\}) \geq 1 - \varepsilon.$$

[10] For the same reason and in contrast to other likelihood-ratio tests in the paper, the test does not check, at time 0, whether or not the law P belongs to the paradigm.

2. For every strategy $\zeta \in \Delta(\mathbb{N} \times \Delta(\Omega))$ there is a law $P_\zeta \in \Lambda$ such that

$$(P_\zeta \otimes \zeta)\{(\omega, (d, P)) : T_{\mu,\varepsilon}(d, \omega, P) = 1\} \leq \varepsilon.$$

Given a testable paradigm Λ, Theorem 1 guarantees we can find a prior $\mu \in \Delta(\overline{\Lambda})$ such that the distance $\|Q_\mu - P\|$ is arbitrarily close to 1 with respect to every law P in the paradigm. Thus, for each ε, there exists a prior that satisfies the assumption of Theorem 4.

The test $T_{\mu,\varepsilon}$ does not reject true experts with probability $1 - \varepsilon$. For every law P in the paradigm there is a deadline d_P such that a true expert who reports a deadline $d \geq d_P$ and then predicts according to P will pass the test with probability greater than $1 - \varepsilon$. The test is also ε-nonmanipulable: For every strategy ζ there is a law in the paradigm under which the forecaster will pass the test with at most ε probability.

The result provides additional justification for the likelihood-ratio test as a methodology for screening forecasters. Compared to other results in the paper, such a justification does not rest on testing off-path, counterfactual, predictions.

Theorem 4 is in contrast with the result of Shmaya (2008), which casted doubts on the possibility of testing strategic forecasters without relying on off-path predictions. Shmaya (2008) considers prequential tests that do not reject the truth with respect to the unrestricted paradigm $\Delta(\Omega)$ and shows that any such test must be manipulable, even if the test is not finite and is allowed to return a decision at infinity. Shmaya's result implies that if no restrictions are imposed on the paradigm, then it is impossible to test strategic forecasters using tests which do not rely on counterfactual predictions. As shown by Theorem 4, both goals are compatible once we consider finite tests and testable paradigms.[11]

7 Identifiable and Maximal Paradigms

In this section, we present and study two representative families of testable paradigms. We first consider *identifiable* paradigms, i.e. paradigms where the true law can be inferred from the data in the long run. We then consider *maximal* paradigms. We show

[11]Nonmanipulable and prequential, but not finite, tests are provided by Stewart (2011), Hu and Shmaya (2013), Sandroni and Shmaya (2014) and Feinberg and Lambert (2015). Fortnow and Vohra (2009) provide a prequential test that is non-manipulable in the presence of a computational constraint on the forecaster.

that any well-behaved testable paradigm is obtained by enlarging some identifiable paradigm. In addition, any testable paradigm can be enlarged to a maximal paradigm.

7.1 Testing and Identification

It is plain that the class of deterministic laws is testable. An uninformed forecaster who is asked to perfectly predict the first n realizations can guarantee only a vanishing probability of not being contradicted by the data. A similar intuition applies to the class of i.i.d. distributions. In this case, an expert must be able to predict, *ex-ante* and with almost perfect precision, what the long run empirical frequency of each outcome will be. The same logic applies to other paradigms of interest for which suitable generalizations of the law of large numbers hold (e.g. the class of all stationary ergodic distributions). The idea common to these examples is that the true law can be eventually inferred from the data. The next definition formalizes this concept.

Definition 9 *A paradigm Λ is* identifiable *if there exists a measurable map $f : \Omega \to \Lambda$ such that $P(\{\omega : f(\omega) = P\}) = 1$ for every $P \in \Lambda$.*

The same notion of identifiability has been studied, in different contexts, by Doob (1949), Blackwell (1980), Weizsäcker (1996) and Al-Najjar and Shmaya (2016) among others. There are several reasons for studying identifiable paradigms more in depth. Notice that experts are *more* relevant when the paradigm is *not* identifiable, because in this case the knowledge held by the expert goes beyond what the tester could infer on his own from the data. It is therefore important to know to what extent testability confines the tester to identifiable paradigms. In addition, as discussed in the previous paragraph, many examples of testable paradigms of practical interest are also identifiable. It is not clear, *a priori*, what is the gap between identifiability and testability. Theorem 5 below clarifies the relationship between the two properties.

The next result is a preliminary observation.

Remark 1 *Any infinite, identifiable paradigm is testable. However, there exist testable and non-identifiable paradigms.*

Hence, any sufficiently rich and identifiable paradigm is testable. However, identifiability is not a necessary condition for testability.[12] To gain an intuition, consider the

[12]Indeed, the testable paradigm presented by Olszewski and Sandroni (2009a) and Al-Najjar, Sandroni, Smorodinsky and Weinstein (2010) are not identifiable.

class $\Lambda_1 = \{\delta_\omega : \omega \in \Omega\}$ of all deterministic laws and now add a mixture $P = \frac{1}{2}\delta_{\omega'} + \frac{1}{2}\delta_{\omega''}$ between two distinct degenerate distributions. The resulting set $\Lambda = \Lambda_1 \cup \{P\}$ is testable but no longer identifiable, since observing the sequence ω' does not reveal whether the data was generated by the mixture P or by the deterministic law $\delta_{\omega'}$.

The next result shows that any suitably well-behaved testable paradigm is obtained by enlarging some rich (i.e. uncountable) identifiable subclass of distributions. Hence, identifiable paradigms form a *minimal* class of testable paradigms.

Theorem 5 *Let Λ be a paradigm for which there exists a prior $\mu \in \Delta(\Lambda)$ such that $\|Q_\mu - P\| = 1$ for every $P \in \Lambda$. Then there exists a paradigm $\tilde{\Lambda} \subseteq \Lambda$ that is identifiable and uncountable. In addition, if Λ is closed then $\tilde{\Lambda}$ can be chosen to be homeomorphic to the class of all deterministic laws.*

The result assumes that Λ is testable and well-behaved. By Theorem 1, testability is equivalent to the existence of a sequence of priors (μ_n) where the distance between each law Q_{μ_n} and the paradigm goes to 1 as n goes to infinity. Theorem 5 assumes that such sequence can be replaced by a single prior $\mu \in \Delta(\Lambda)$. This assumption can be interpreted as a form of continuity with respect to the total-variation distance.[13] The result also shows that when Λ is closed, then $\tilde{\Lambda}$ can be chosen to be topologically equivalent to the paradigm of all degenerate distributions.

7.2 Maximal Paradigms

Any theory about the data generating process, if incorrect, exposes the tester to the risk of rejecting informed experts. Following this motivation, Feinberg and Stewart (2008), Olszewski and Sandroni (2009), Stewart (2011) and Feinberg and Lambert (2015) develop nonmanipulable tests that do not reject true expert except for a topologically small set of distributions. Relatedly, Olszewski (2015) posed the question of which testable paradigms are *maximal*, in the sense of not being included in any other testable paradigm. The next result provides an answer to this open question. Say that Λ is *ε-testable* if it admits a test that passes the truth with probability $1 - \varepsilon$ and is ε-nonmanipulable.

[13] The paradigm of Al-Najjar, Sandroni, Smorodinsky and Weinstein (2010), all the examples mentioned at the beginning of this section satisfy this property.

Theorem 6 *Let $\varepsilon \in (0,1)$ Given a law $Q \in \Delta(\Omega)$ the paradigm*

$$\Lambda_Q^\varepsilon = \{P \in \Delta(\Omega) : \|Q - P\| > 1 - \varepsilon\}$$

is ε-testable. In addition, it is not included in any testable paradigm.

The set Λ_Q^ε is obtained by fixing a distribution Q and considering all laws which are sufficiently far from P. The resulting class Λ_Q^ε is not included in any other testable paradigm (indeed, as shown in the proof, Λ_Q^ε is included in any δ-testable paradigm, for δ sufficiently small). By Theorem 2, given paradigm Λ_Q^ε, forecasters can be tested using a simple likelihood-ratio test with respect to the benchmark Q.

The adoption of maximal paradigms has two effects. On the one hand, maximal paradigms reduce the risk of accidentally rejecting true experts. On the other hand, they make nonmanipulability a weaker concept. This is because the assertion that uninformed forecasters are screened out by non-manipulable tests rests on the assumption that uninformed agents evaluate the odds of passing the test according to the worst-case scenario with respect to the paradigm Λ (as discussed in Section 2.3). Such an assumption becomes stronger as the paradigm gets larger.

The next result shows that this tension disappears if maximal paradigms are obtained by enlarging an already testable paradigm.

Theorem 7 *Let Λ be a testable paradigm. Then, for every $\varepsilon > 0$ there exist a prior $\mu \in \Delta(\overline{\Lambda})$ such that $\Lambda \subseteq \Lambda_{Q_\mu}^\varepsilon$ and a test T such that:*

1. *For every law $P \in \Lambda_{Q_\mu}^\varepsilon$, $E_P[T(\cdot, P)] \geq 1 - \varepsilon$; and*
2. *For every strategy ζ there exists a law $P_\zeta \in \Lambda$ such that $E_{P_\zeta \otimes \zeta}[T] \leq \varepsilon$.*

The first part of the result shows that any testable paradigm Λ can be extended to a maximal paradigm $\Lambda_{Q_\mu}^\varepsilon$. The second part of the result shows that there exists a test T that is likely to pass a true expert with respect to any law in the maximal paradigm $\Lambda_{Q_\mu}^\varepsilon$, and such that for every strategy there is a law *in the original paradigm* Λ under which a strategic forecaster is likely to be rejected. So, a strategic forecaster who evaluates the odds of passing the tests by considering the worst-case scenario with respect to the original paradigm Λ is screened out by the test. The result shows that any testable paradigm can be enlarged to minimize the risk of rejecting true experts and without affecting the deterrent effect on strategic forecasters.

8 Discussion and Extensions

8.1 Markov Processes

The characterization of Theorem 1 is now applied to the paradigm of all Markov distributions, which is shown to be testable.

Each Markov law is described by a transition probability $\pi : X \to \Delta(X)$ and an initial probability $\rho \in \Delta(X)$. Every pair (ρ, π) induces a Markov distribution $P_{\rho,\pi} \in \Delta(\Omega)$. We now define the prior of the Bayesian outside observer. Fix two distinct outcomes x and y. Given $\alpha \in [0,1]$, let π_α be the transition probability defined as

$$\pi_\alpha(x)(x) = \alpha,\ \pi_\alpha(x)(y) = 1-\alpha \text{ and } \pi_\alpha(z)(x) = 1 \text{ for all } z \neq x$$

Thus, if the current outcome is x then the process remains at x with probability α and moves to y with probability $1-\alpha$. If the current outcome is any z other than x, then the process returns to x almost surely. To simplify the notation, let P_α be the Markov distribution where the two outcomes x and y have initial probability $\frac{1}{2}$ and the transition probability is π_α. Consider a Bayesian outside observer who is uncertain about the transition probability of the process and believes the true law to be P_α for some α. By taking α to be uniformly distributed in $(0,1)$, we obtain (implicitly) a prior μ defined on the set of Markov distributions such that the resulting law Q_μ satisfies

$$Q_\mu(E) = \int_0^1 P_\alpha(E)\, d\alpha$$

for every event E.

Theorem 8 *The prior μ satisfies $\|Q_\mu - P_{\pi,\rho}\| = 1$ for all Markov $P_{\pi,\rho}$.*

The result is obtained by applying standard asymptotic results for Markov processes. Theorems 1 and 2 allow then to conclude that the paradigm is testable by means of a likelihood-ratio test with respect to the law Q_μ.

8.2 Non-asymptotic Tests

We now consider the case where at most n observations are available to the tester. The main conclusion is that most of the results obtained for general tests (Theorems 1, 2 and 4) can be adapted to non-asymptotic tests.

Call a paradigm *ε-testable in n periods* if it admits a test T such that $T(\cdot, P)$ is $\mathcal{F}_n$-measurable for every P, does not reject the truth with probability $1-\varepsilon$, and is ε-nonmanipulable. The next result mirrors the characterization of Theorem 1. Given n and two laws P and Q, we denote (with a slight abuse of notation) by $\|Q-P\|_n$ the distance $\max_{E\in\mathcal{F}_n} |Q(E)-P(E)|$.

Theorem 9 *Let Λ be a paradigm. If there exists a prior $\mu \in \Delta\left(\overline{\Lambda}\right)$ with the property that $\|Q_\mu - P\|_n > 1-\varepsilon$ for every $P \in \Lambda$, then Λ is ε-testable in n periods. Conversely, if Λ is ε-testable in n periods then there is a prior $\mu \in \Delta\left(\overline{\Lambda}\right)$ such that $\|Q_\mu - P\|_n > 1-2\varepsilon$ for every $P \in \Lambda$.*

Hence, similarly to Theorem 1, testability in n periods is equivalent to a high distance between the law Q_μ induced by the prior and each law in the paradigm. The main difference between the two results is that in Theorem 9 the distance between laws is now computed with respect to events occurring before time n.

The next result shows that for paradigms that are ε-testable in n periods, restricting the attention to to prequential likelihood-ratio tests is loss of generality for ε sufficiently small.

Theorem 10 *Let Λ be a paradigm. If $\mu \in \Delta\left(\overline{\Lambda}\right)$ satisfies $\|Q_\mu - P\|_n > 1-\varepsilon^2/(1+\varepsilon)$ for every $P \in \Lambda$, then the test*

$$T(\omega, P) = \begin{cases} 1 & \text{if } P(\omega^n) > \frac{1}{\varepsilon} Q_\mu(\omega^n) \\ 0 & \text{otherwise} \end{cases}$$

is prequential, does not reject the truth with probability $1-\varepsilon$ and is ε-nonmanipulable.

Fix $\varepsilon > 0$ and let $\delta = \varepsilon^2/(1+\varepsilon)$. The result shows that if the paradigm is δ-testable in n periods, then there exists a *prequential* test that does not reject the truth with probability $1-\varepsilon$ and is ε nonmanipulable.

8.3 Maxmin and Strategic Forecasters

Consider a strategic forecaster producing his forecasts according to a strategy ζ. As discussed in Section 4, a strategic but uninformed forecaster who adopts the Gilboa-Schmeidler maxmin criterion will evaluate the strategy according to

$$\inf_{P\in C} E_{P\otimes\zeta}\left[wT + l(1-T)\right]$$

where C is a set of laws. So far, we have assumed C to be equal to the paradigm Λ under consideration. However, an uninformed forecaster may adopt a less conservative decision making criterion. To this end we fix a distance for the weak*-topology on $\Delta(\Omega)$, and for every law $P \in \Delta(\Omega)$ denote by $B_\delta(P)$ the open ball of radius δ around P. A natural specification for the set C of of laws takes the form

$$C = B_\delta(P_o) \cap \Lambda \text{ for some } P_o \in \Lambda. \tag{8}$$

Under this specification, the uninformed forecaster evaluates each strategy by considering the worst-case expected payoff with respect to laws that are within a distance δ from a reference measure P_o. Similar specifications appear in robust statistics (Huber, 1981) and economics (Bergemann and Schlag, 2011, and Babaioff, Blumrosen, Lambert and Reingold, 2010). We will not assume that P_o coincides with the correct law generating the data nor that P_o is known to the tester.

We now consider the problem of screening between informed and uninformed under the specification (8).

Definition 10 *A paradigm Λ is* uniformly testable with precision δ *if for every $\varepsilon > 0$ there exists a finite test T such that:*

1. *T does not reject the truth with probability $1 - \varepsilon$; and*
2. *For every strategy ζ and every $P_o \in \Lambda$ there exists a law $P_\zeta \in \Lambda \cap B_\delta(P_o)$ such that $E_{P_\zeta \otimes \zeta}[T] \leq \varepsilon$.*

Thus, the test passes a true expert with high probability. In addition, for every strategy ζ there is a law P_ζ in the paradigm under which rejection is likely. In addition, for every reference law P_o, the measure P_ζ can be chosen to belong to $\Lambda \cap B_\delta(P_o)$. Hence, the test guarantees that the value an uninformed forecaster can expect from participating in the test is negative whenever ε is sufficiently small. So, the test can screen between the two types of forecasters.

The next theorem provides a simple sufficient condition for a paradigm to be uniformly testable.

Theorem 11 *Let $\mu \in \Delta(\overline{\Lambda})$ be a prior with support $\overline{\Lambda}$ such that Q_μ satisfies $\|Q_\mu - P\| = 1$ for every P in Λ. Then Λ is uniformly testable with precision δ for every $\delta > 0$.*

Compared to the characterization of Theorem 1, the prior μ is now required to have support equal to the closure of the paradigm. An exact characterization of uniformly testable paradigms is provided in Appendix B.

A Appendix

A.1 Preliminaries

The space of paths Ω is endowed with the product topology. Hence, a function that is $\mathcal{F}_n$-measurable for some n is also continuous. This implies that for every finite test T and any law $P \in \Delta(\Omega)$ the function $Q \mapsto E_Q[T(\cdot, P)]$, $Q \in \Delta(\Omega)$, is continuous. We will denote by $\mathcal{H}_n$ the set of histories ω^n of length n.

Recall that the space $\Delta(\Delta(\Omega))$ is endowed with the weak* topology. As proved in Phelps (2001) (Proposition 1.1), the function $\mu \mapsto Q_\mu$ assigning to each prior $\mu \in \Delta(\Delta(\Omega))$ its barycenter Q_μ is continuous. In particular, given a continuous function $\psi : \Omega \to \mathbb{R}$, the map $\mu \mapsto \int_\Omega \psi(\omega)\, dQ_\mu(\omega)$, $\mu \in \Delta(\Delta(\Omega))$, is continuous. In addition, Q_μ satisfies $\int_\Omega \psi(\omega)\, dQ_\mu(\omega) = \int_{\Delta(\Omega)} \left(\int_\Omega \psi(\omega)\, dQ(\omega)\right) d\mu(Q)$ for every bounded measurable function ψ. Given a measurable subset Γ of $\Delta(\Omega)$, denote by $\Delta(\Gamma)$ the set of probability measures $P \in \Delta(\Omega)$ assigning probability 1 to Γ. The space $\Delta\left(\overline{\Gamma}\right)$ is compact by the Banach-Alaoglu theorem (see Aliprantis and Border (2006, Chapter 16)).

Lemma 1 *Let T be a finite test. For every strategy ζ the function $P \mapsto E_{P\otimes\zeta}[T]$, $P \in \Delta(\Omega)$, is continuous.*

Proof. Let (ω_k) be a sequence in Ω converging to a path ω. Given a law P, the function $T(\cdot, P)$ is continuous. So, $T(\omega_k, P) \to T(\omega, P)$ as $k \to \infty$. Given a strategy ζ, Lebesgue's convergence theorem implies $E_\zeta[T(\omega_k, \cdot)] \to E_\zeta[T(\omega, \cdot)]$ as $k \to \infty$. Hence, for every strategy ζ the map $\omega \mapsto E_\zeta[T(\omega, \cdot)]$, $\omega \in \Omega$, is continuous. Fubini's Theorem implies $E_{P\otimes\zeta}[T] = \int_\Omega E_\zeta[T(\omega, \cdot)]\, dP(\omega)$. Therefore, for each P, $\int_\Omega E_\zeta[T(\omega, \cdot)]\, dP(\omega)$ is the expectation with respect to P of a continuous function. Hence, it follows from the definition of weak* topology that the map $P \mapsto E_{P\otimes\zeta}[T]$, $P \in \Delta(\Omega)$, is continuous. ■

A.2 Proofs of Results of Sections 3 and 4

Proof of Theorems 1 and 2. The first part of the proof shows the necessity part of Theorem 1. The second half of the proof establishes Theorem 2 and, therefore, the sufficiency part of Theorem 1.

Assume that Λ is testable. Fix $\varepsilon > 0$ and let T be a test that satisfies the conditions of Definition 5. Given a measure $P \in \Delta(\Omega)$ and a strategy ζ, let $V(P, \zeta) = E_{P \otimes \zeta}[T]$. The map V is affine in each argument and for each strategy ζ the map $V(\cdot, \zeta)$ is continuous by Lemma 1. Since T is ε-nonmanipulable then

$$\sup_{\zeta \in \Delta(\Delta(\Omega))} \inf_{P \in \Lambda} V(P, \zeta) \leq \varepsilon. \tag{9}$$

Let $\Delta_o(\Lambda) \subseteq \Delta(\Lambda)$ be the subset of priors on Λ with finite support. We have

$$\sup_{\zeta \in \Delta(\Delta(\Omega))} \inf_{P \in \Lambda} V(P, \zeta) = \sup_{\zeta \in \Delta(\Delta(\Omega))} \inf_{\mu \in \Delta_o(\Lambda)} V(Q_\mu, \zeta) = \sup_{\zeta \in \Delta(\Delta(\Omega))} \min_{\mu \in \Delta\left(\overline{\Lambda}\right)} V(Q_\mu, \zeta). \tag{10}$$

The first equality follows immediately from the definition of Q_μ and the affinity of $V(\cdot, \zeta)$. The second equality follows from the continuity of the map $\mu \mapsto V(Q_\mu, \zeta)$, $\mu \in \Delta(\Delta(\Omega))$, together with the fact that $\Delta_o(\Lambda)$ is dense in $\Delta\left(\overline{\Lambda}\right)$ (as implied by Aliprantis and Border (2006, Theorem 15.10)) and that $\Delta\left(\overline{\Lambda}\right)$ is compact.

The space $\Delta\left(\overline{\Lambda}\right)$ is compact and convex and for every ζ the map $\mu \mapsto V(Q_\mu, \zeta)$, $\mu \in \Delta(\Delta(\Omega))$, is continuous and affine. In addition, $\Delta(\Delta(\Omega))$ is convex and for every μ, the map $V(Q_\mu, \cdot)$ is affine. We can therefore apply Fan's Minmax Theorem (Fan (1953)) to obtain the equality

$$\sup_{\zeta \in \Delta(\Delta(\Omega))} \min_{\mu \in \Delta\left(\overline{\Lambda}\right)} V(Q_\mu, \zeta) = \min_{\mu \in \Delta\left(\overline{\Lambda}\right)} \sup_{\zeta \in \Delta(\Delta(\Omega))} V(Q_\mu, \zeta). \tag{11}$$

For every μ, the function V satisfies $V(Q_\mu, \zeta) = \int_{\Delta(\Omega)} E_{Q_\mu}[T(\cdot, P)]\, d\zeta(P)$ by Fubini's theorem. So, $\sup_{\zeta \in \Delta(\Delta(\Omega))} V(Q_\mu, \zeta) = \sup_{P \in \Delta(\Omega)} V(Q_\mu, \delta_P)$. Hence the right-hand side of (11) can be written as

$$\min_{\mu \in \Delta\left(\overline{\Lambda}\right)} \sup_{\zeta \in \Delta(\Delta(\Omega))} V(Q_\mu, \zeta) = \min_{\mu \in \Delta\left(\overline{\Lambda}\right)} \sup_{P \in \Delta(\Omega)} V(Q_\mu, \delta_P) = \min_{\mu \in \Delta\left(\overline{\Lambda}\right)} \sup_{P \in \Delta(\Omega)} E_{Q_\mu}[T(\cdot, P)]. \tag{12}$$

Taken together, (9), (10) (11) and (12) prove the existence of a prior $\mu \in \Delta\left(\overline{\Lambda}\right)$ such

that

$$\sup_{\zeta\in\Delta(\Delta(\Omega))}\inf_{P\in\Lambda}V\left(P,\zeta\right)=\sup_{P\in\Delta(\Omega)}E_{Q_{\mu}}\left[T\left(\cdot,P\right)\right]\leq\varepsilon.$$

Because the test does not reject the truth with probability $1-\varepsilon$, it follows that

$$E_{P}\left[T\left(\cdot,P\right)\right]-E_{Q_{\mu}}\left[T\left(\cdot,P\right)\right]\geq1-2\varepsilon\text{ for all }P\in\Lambda. \tag{13}$$

As shown by Lemmas 1 and 2 in Shiryaev (2016, Chapter 8), the (normalized) total variation distance $\left\|Q_{\mu}-P\right\|$ satisfies

$$\left\|Q_{\mu}-P\right\|=\sup_{\phi}\left|\int_{\Omega}\phi dQ_{\mu}-\int_{\Omega}\phi dP\right|$$

where the supremum is taken over all measurable functions $\phi:\Omega\rightarrow[0,1]$. By letting $\phi=T\left(\cdot,P\right)$, it follows from (13) that $\left\|Q_{\mu}-P\right\|\geq1-2\varepsilon$. Thus, $\left\|Q_{\mu}-P\right\|\geq1-2\varepsilon$ for every $P\in\Lambda$. Since ε is arbitrary, the first part of the proof is concluded.

Consider a prior $\mu\in\Delta\left(\overline{\Lambda}\right)$ such that $\left\|Q_{\mu}-P\right\|>1-\varepsilon$ for all $P\in\Lambda$. Fix a measure $P\in\Lambda$. For any n,

$$\max_{E\in\mathcal{F}_{n}}Q_{\mu}\left(E\right)-P\left(E\right)=\max_{E\in\mathcal{F}_{n}}\left|Q_{\mu}\left(E\right)-P\left(E\right)\right|.$$

As shown in Halmos (1950, 13D), $\max_{E\in\mathcal{F}_{n}}Q_{\mu}\left(E\right)-P\left(E\right)\uparrow\left\|Q_{\mu}-P\right\|$ as $n\uparrow\infty$. Therefore, we can conclude that for each $P\in\Lambda$ the number

$$n_{P}=\min\left\{n:\max_{E\in\mathcal{F}_{n}}Q_{\mu}\left(E\right)-P\left(E\right)>1-\varepsilon\right\} \tag{14}$$

is well defined. Consider now the test

$$T\left(\omega,P\right)=\begin{cases}1 & \text{if }P\in\Lambda\text{ and }P\left(\omega^{n_{P}}\right)>Q_{\mu}\left(\omega^{n_{P}}\right)\\ 0 & \text{otherwise}\end{cases}$$

We now prove that T is measurable. First we show that for every $k\in\mathbb{N}$ the set $\left\{P\in\Lambda:n_{P}=k\right\}$ is measurable. For every n and every $E\in\mathcal{F}_{n}$ the function $P\mapsto P\left(E\right),\ P\in\Delta\left(\Omega\right)$, is continuous. Because $\mathcal{F}_{n}$ is finite, it follows that $\varphi_{n}:P\mapsto\max_{E\in\mathcal{F}_{n}}Q_{\mu}\left(E\right)-P\left(E\right),\ P\in\Delta\left(\Omega\right)$, is measurable. Since Λ is measurable the restriction of φ_{n} on Λ is also measurable. The set $\left\{P\in\Lambda:n_{P}=k\right\}$ can be written as

$\{P \in \Lambda : \varphi_k > 1-\varepsilon\}$ if $k = 1$, or as the intersection

$$\bigcap_{1 \leq n < k} \{P \in \Lambda : \varphi_n \leq 1-\varepsilon\} \cap \{P \in \Lambda : \varphi_k > 1-\varepsilon\}$$

if $k > 1$. Hence $\{P \in \Lambda : n_P = k\}$ is measurable. For each path ω, the function $T(\omega, \cdot)$ is measurable: For each n, the set $\{P \in \Delta(\Omega) : T(\omega, P) = 1\}$ is given by the union over $k > 1$ of all sets of the form

$$\left\{P \in \Delta(\Omega) : P\left(\omega^k\right) - Q_\mu\left(\omega^k\right) > 0\right\} \cap \{P \in \Lambda : n_P = k\}.$$

It follows that $T(\omega, \cdot)$ is measurable. For each $\omega \in \Omega$ and $P \in \Delta(\Omega)$, the function $T(\cdot, P)$ is continuous and $T(\omega, \cdot)$ is measurable. That is, T is a Carathéodory functions. It follows then from Lemma 4.51 in Aliprantis and Border (2016) that T is measurable.

We now show that $P(\{\omega : T(\omega, P) = 1\}) > 1-\varepsilon$ and $Q_\mu(\{\omega : T(\omega, P) = 1\}) < \varepsilon$ for each P. The proof follows Lehmann and Romano (2006, Chapter 16). If $P \notin \Lambda$ the result is obvious. So let $P \in \Lambda$, and denote be A^P the set $\{\omega : P(\omega^{n_P}) > Q_\mu(\omega^{n_P})\}$. Let $\mathcal{H}_{n_P}$ be the set of all histories of length n_P. Then for every $E \in \mathcal{F}_{n_P}$ we have

$$\begin{aligned} P(E) - Q_\mu(E) &= \sum_{\omega^{n_P} \in \mathcal{H}_{n_P} : \omega^{n_P} \subseteq E} P(\omega^{n_P}) - Q_\mu(\omega^{n_P}) \\ &\leq \sum_{\omega^{n_P} \in \mathcal{H}_{n_P} : \omega^{n_P} \subseteq E \cap A^P} P(\omega^{n_P}) - Q_\mu(\omega^{n_P}) \\ &\leq \sum_{\omega^{n_P} \in \mathcal{H}_{n_P} : \omega^{n_P} \subseteq A^P} P(\omega^{n_P}) - Q_\mu(\omega^{n_P}). \end{aligned}$$

Therefore $P\left(A^P\right) - Q_\mu\left(A^P\right) = \max_{E \in \mathcal{F}_{n_P}} P(E) - Q_\mu(E) > 1-\varepsilon$. So $P\left(A^P\right) > 1-\varepsilon$ (in particular, the test T does not reject the truth with probability $1-\varepsilon$) and $Q_\mu\left(A^P\right) < \varepsilon$. We can now show that T is ε-nonmanipulable. For every strategy ζ, we have

$$V(Q_\mu, \zeta) = \int_{\Delta(\Omega)} Q_\mu\left(A^P\right) d\zeta(P) < \varepsilon.$$

Using again the fact that $\mu \mapsto V(Q_\mu, \zeta)$, $\mu \in \Delta(\Delta(\Omega))$, is continuous and $\Delta_o(\Lambda)$ is dense in $\Delta\left(\overline{\Lambda}\right)$, we can find a prior $\mu_\zeta \in \Delta_o(\Lambda)$ such that

$$V\left(Q_{\mu_\zeta}, \zeta\right) = \sum_{P \in \Lambda} \mu_\zeta(P) V(P, \zeta) < \varepsilon$$

Hence, there must exists some law $P_\zeta \in \Lambda$ in the support of μ_ζ such that $V(P_\zeta, \zeta) < \varepsilon$. Because ε is arbitrary, we conclude that Λ is testable. ■

Proof of Corollary 1. As shown in Phelps (2001, Proposition 1.2) a law P belongs to the weak*-closed convex hull of Λ if and only if there exists a prior $\mu \in \Delta\left(\overline{\Lambda}\right)$ such that $P = Q_\mu$. The result now follows immediately from Theorem 1. ■

A.3 Proof of Theorem 3

The next result is a version of the Neyman-Pearson lemma. The standard proof parallels the proof of Theorem 3.2.1 in Lehmann and Romano, (2006) and is therefore omitted.

Theorem 12 (Neyman-Pearson Lemma) *Let $P_0, P_1 \in \Delta(\Omega)$. Given $n \in \mathbb{N}$ and $\alpha \in [0,1]$, let Φ be the set of $\mathcal{F}_n$-measurable functions $\phi : \Omega \to [0,1]$ that satisfy $E_{P_0}[\phi] \geq \alpha$. Let*

$$\lambda = \sup\left\{k \in \mathbb{R} : P_0\left(\left\{\omega : P_0\left(\omega^n\right) \geq kP_1\left(\omega^n\right)\right\}\right) \geq \alpha\right\}$$

and, letting $0 \cdot \infty = 0$, define

$$\begin{aligned} \delta &= P_0\left(\left\{\omega : P_0\left(\omega^n\right) > \lambda P_1\left(\omega^n\right)\right\}\right) \\ \gamma &= P_0\left(\left\{\omega : P_0\left(\omega^n\right) = \lambda P_1\left(\omega^n\right)\right\}\right) \end{aligned}$$

The function

$$\phi^*(\omega) = \begin{cases} 1 & \text{if } P_0\left(\omega^n\right) > \lambda P_1\left(\omega^n\right) \\ \frac{\alpha-\delta}{\gamma} & \text{if } P_0\left(\omega^n\right) = \lambda P_1\left(\omega^n\right) \text{ and } \gamma > 0 \\ 0 & \text{otherwise} \end{cases}$$

is a solution to $\min_{\phi\in\Phi} E_{P_1}[\phi]$.

Proof of Theorem 3. Fix a paradigm Λ, testing times (n_P) and a probability $\alpha \in [0,1]$. Denote by $\mathcal{T}$ the class of finite tests that are bounded by (n_P) and do not reject the truth with probability α.

For every $P \in \Lambda$, let Φ_P be the set of $\mathcal{F}_{n_P}$-measurable functions $\phi : \Omega \to [0,1]$ that satisfy $E_P[\phi] \geq \alpha$. Define the function $f : \Delta\left(\overline{\Lambda}\right) \to \mathbb{R}$ as

$$f(\mu) = \sup_{P\in\Lambda} \min_{\phi\in\Phi_P} E_{Q_\mu}[\phi].$$

The function f is lower-semicontinous: Fix $P \in \Lambda$. The set Φ_P can be identified with a subset of $[0,1]^m$, where m is the cardinality of the set of histories of length n_P. It is then immediate to verify that Φ_P is compact. It then follows from the theorem of the maximum that the map $Q \mapsto \min_{\phi \in \Phi_P} E_{Q_\mu}[\phi]$, $Q \in \Delta(\Omega)$, is continuous. Thus, the continuity of $\mu \mapsto Q_\mu$, $\mu \in \Delta(\Delta(\Omega))$ implies that the map $\mu \mapsto \min_{\phi \in \Phi_P} E_{Q_\mu}[\phi]$, $\mu \in \Delta(\Delta(\Omega))$ is a composition of continuous functions. Thus, f is a supremum of continuous functions. Hence f is lower-semicontinuous and so attains a minimum on $\Delta\left(\overline{\Lambda}\right)$. Let μ^* be a prior which minimizes f.

Denote by ϕ_P^* the test obtained by applying the Neyman-Pearson lemma when setting $P_0 = P$, $P_1 = Q_{\mu^*}$ and $n = n_P$ in the statement of Theorem 12. Denote also by λ_P, δ_P and γ_P the corresponding quantities. Let T^* be the test defined as

$$T^*(\omega, P) = \begin{cases} \phi_P^*(\omega) & \text{if } P \in \Lambda \\ 0 & \text{if } P \notin \Lambda. \end{cases}$$

We now show that T^* is a well-defined test belonging to $\mathcal{T}$. By definition, the test is finite and does not reject the truth with probability α. It remains to show it is measurable. By Lemma 4.51 in Aliprantis and Border (2016), it is enough to prove that $T(\omega, \cdot)$ is measurable for every ω. We first show that the map $P \mapsto \lambda_P$, $P \in \Lambda$, mapping each measure to the corresponding threshold $\lambda_P \in [0, \infty]$ in the likelihood-ratio test, is measurable. For every $k \in \mathbb{R}$ let

$$\Gamma_k = \left\{P \in \Lambda : P\left(\left\{\omega : P\left(\omega^{n_P}\right) \geq k Q_{\mu^*}\left(\omega^{n_P}\right)\right\}\right) \geq \alpha\right\}.$$

Notice that Γ_k can be written as

$$\bigcup_{m \in \mathbb{N}} \left(\left\{P \in \Lambda : n_P = m\right\} \cap \left\{P \in \Lambda : P\left(\left\{\omega : P\left(\omega^m\right) \geq k Q_{\mu^*}\left(\omega^m\right)\right\}\right) \geq \alpha\right\}\right)$$

Each set $\{P \in \Lambda : n_P = m\}$ is measurable. For each ω^m the function $P \mapsto P(\omega^m)$, $P \in \Delta(\Omega)$, is continuous. So, for each history ω^m the set

$$\Upsilon_{\omega^m} = \left\{P \in \Lambda : P\left(\omega^m\right) \geq k Q_{\mu^*}\left(\omega^m\right)\right\}$$

is measurable. Let $1_{\Upsilon_{\omega^m}}$ be the indicator function of Υ_{ω^m} and notice that

$$P\left(\left\{\omega : P\left(\omega^m\right) \geq kQ_{\mu^*}\left(\omega^m\right)\right\}\right) = \sum_{\omega^m \in \mathcal{H}_m} P\left(\omega^m\right) 1_{\Upsilon_{\omega^m}}\left(P\right),$$

where the latter is a measurable function of P. It then follows that each set of the form

$$\left\{P \in \Lambda : P\left(\left\{\omega : P\left(\omega^m\right) \geq kQ_{\mu^*}\left(\omega^m\right)\right\}\right) \geq \alpha\right\}$$

is measurable. Thus, Γ_k is measurable. This in turn yields that for each k the function $P \mapsto k1_{\Gamma_k}(P)$ is measurable. Notice that $\lambda_P = \sup_{k\in\mathbb{Q}} k1_{\Gamma_k}(P)$ for every P. Thus, we can conclude that the function $P \mapsto \lambda_P$ (mapping $\Delta(\Omega)$ to $\mathbb{R} \cup \{\infty\}$) is measurable. Now fix a path ω. An argument analogous to that one used to prove the measurability of the set Γ_k shows that $\{P \in \Lambda : P(\omega^{n_P}) > \lambda_P Q_{\mu^*}(\omega^{n_P})\}$ and $\{P \in \Lambda : P(\omega^{n_P}) = \lambda_P Q_{\mu^*}(\omega^{n_P})\}$ are measurable and that δ_P and γ_P are measurable functions of P. It is then routine to verify that $T(\omega, \cdot)$ is measurable. We can therefore conclude that T is a well defined test belonging to $\mathcal{T}$.

We now show that T^* is a least manipulable test in the class $\mathcal{T}$. Let $T \in \mathcal{T}$. As in the proof of Theorems 1 and 2, given any test $T \in \mathcal{T}$ we can apply Fan's minmax theorem to conclude

$$\sup_{\zeta\in\Delta(\Delta(\Omega))} \inf_{P\in\Lambda} E_{P\otimes\zeta}[T] = \min_{\mu\in\Delta\left(\overline{\Lambda}\right)} \sup_{P\in\Delta(\Omega)} E_{Q_\mu}\left[T\left(\cdot, P\right)\right]. \tag{15}$$

It is without loss of generality to assume that $T(\omega, P) = 0$ for every ω and $P \notin \Lambda$. So, the expression can be simplified to

$$\sup_{\zeta\in\Delta(\Lambda)} \inf_{P\in\Lambda} E_{P\otimes\zeta}[T] = \min_{\mu\in\Delta\left(\overline{\Lambda}\right)} \sup_{P\in\Lambda} E_{Q_\mu}\left[T\left(\cdot, P\right)\right].$$

The test T is finite and does not reject the truth with probability α. So, it satisfies $T(\cdot, P) \in \Phi_P$ for every $P \in \Lambda$. Thus,

$$\begin{aligned}
\min_{\mu\in\Delta\left(\overline{\Lambda}\right)} \sup_{P\in\Lambda} E_{Q_\mu}\left[T\left(\cdot, P\right)\right] &\geq \min_{\mu\in\Delta\left(\overline{\Lambda}\right)} \sup_{P\in\Lambda} \min_{\phi\in\Phi_P} E_{Q_\mu}\left[\phi\right] \\
&= \min_{\mu\in\Delta\left(\overline{\Lambda}\right)} f\left(\mu\right) \\
&= \sup_{P\in\Lambda} \min_{\phi\in\Phi_P} E_{Q_{\mu^*}}\left[\phi\right].
\end{aligned}$$

The essential idea is that the test T^* has been defined to satisfy

$$E_{Q_{\mu^*}}\left[T^*\left(\cdot,P\right)\right]=\min_{\phi\in\Phi_P}E_{Q_{\mu^*}}\left[\phi\right]$$

for every $P\in\Lambda$. This means that

$$\begin{aligned}\min_{\mu\in\Delta\left(\overline{\Lambda}\right)}\sup_{P\in\Lambda}E_{Q_\mu}\left[T\left(\cdot,P\right)\right] &\geq \sup_{P\in\Lambda}\min_{\phi\in\Phi_P}E_{Q_{\mu^*}}\left[\phi\right]\\ &= \sup_{P\in\Lambda}E_{Q_{\mu^*}}\left[T^*\left(\cdot,P\right)\right]\\ &\geq \min_{\mu\in\Delta\left(\overline{\Lambda}\right)}\sup_{P\in\Lambda}E_{Q_\mu}\left[T^*\left(\cdot,P\right)\right].\end{aligned}$$

By applying (15) to both T and T^* we now obtain

$$\sup_{\zeta\in\Delta(\Lambda)}\inf_{P\in\Lambda}E_{P\otimes\zeta}\left[T\right]\geq\sup_{\zeta\in\Delta(\Lambda)}\inf_{P\in\Lambda}E_{P\otimes\zeta}\left[T^*\right].$$

Hence, T^* is less manipulable than T. ■

A.4 Proof of Theorem 4

In the course of the proof of Theorem 4 we will use the following basic estimates.

Lemma 2 *Let $P,Q\in\Delta\left(\Omega\right)$. For every $n\in\mathbb{N}$ and every $t>0$,*

$$Q\left(\left\{\omega:tP\left(\omega^n\right)\geq Q\left(\omega^n\right)\right\}\right)\leq t \quad \textit{and} \quad Q\left(\left\{\omega:tP\left(\omega^n\right)>Q\left(\omega^n\right)\right\}\right)<t.$$

Proof. Recall that $\mathcal{H}_n$ is the set of histories of length n. We have,

$$\begin{aligned}Q\left(\left\{\omega:tP\left(\omega^n\right)\geq Q\left(\omega^n\right)\right\}\right) &= \sum_{\omega^n\in\mathcal{H}_n:tP(\omega^n)\geq Q(\omega^n)}Q\left(\omega^n\right) \qquad (16)\\ &\leq t\left(\sum_{\omega^n\in\mathcal{H}_n:tP(\omega^n)\geq Q(\omega^n)}P\left(\omega^n\right)\right)\\ &= tP\left(\left\{\omega:tP\left(\omega^n\right)\geq Q\left(\omega^n\right)\right\}\right)\\ &\leq t.\end{aligned}$$

For every history ω^n, if $tP\left(\omega^n\right)>Q\left(\omega^n\right)$ then $P\left(\omega^n\right)>0$. So, similarly to (16), we

have

$$Q\left(\left\{\omega : tP\left(\omega^n\right) > Q\left(\omega^n\right)\right\}\right) < t\left(\sum_{\omega^n \in \mathcal{H}_n : tP(\omega^n) > Q(\omega^n)} P\left(\omega^n\right)\right)$$

so, $Q\left(\left\{\omega : tP\left(\omega^n\right) > Q\left(\omega^n\right)\right\}\right) < t$. ■

The next result relates the total variation distance and likelihood-ratios.

Lemma 3 *Let $P, Q \in \Delta(\Omega)$ satisfy*

$$\max_{E \in \mathcal{F}_{n_o}} |P(E) - Q(E)| > 1 - \varepsilon$$

for some $\varepsilon > 0$ and $n_o \in \mathbb{N}$. Then, for every $n \geq n_o$ and for every $k > 0$,

$$P\left(\left\{\omega : P\left(\omega^n\right) > kQ\left(\omega^n\right)\right\}\right) > 1 - k\varepsilon - \varepsilon.$$

Proof. Because $\mathcal{F}_{n_o} \subseteq \mathcal{F}_n$ for every $n \geq n_o$, then

$$\max_{E \in \mathcal{F}_n} |P(E) - Q(E)| = \max_{E \in \mathcal{F}_n} P(E) - Q(E) > 1 - \varepsilon \text{ for every } n \geq n_o.$$

Now fix $n \geq n_o$. Let E_n be an event in $\mathcal{F}_n$ such that $P(E_n) - Q(E_n) > 1 - \varepsilon$. So $P(E_n) > 1 - \varepsilon$ and $Q(E_n) < \varepsilon$. Let $A_n^+ = \{\omega : Q(\omega^n) > 0\}$. Notice that $A_n^+ \in \mathcal{F}_n$ and for every $A \in \mathcal{F}_n$ we have

$$P\left(A \cap A_n^+\right) = \sum_{\omega^n \in \mathcal{H}_n : \omega^n \subseteq A \cap A_n^+} P\left(\omega^n\right) = \sum_{\omega^n \in \mathcal{H}_n : \omega^n \subseteq A \cap A_n^+} \frac{P\left(\omega^n\right)}{Q\left(\omega^n\right)} Q\left(\omega^n\right)$$

which can be conveniently rewritten as

$$P\left(A \cap A_n^+\right) = \int_{A \cap A_n^+} \frac{P\left(\omega^n\right)}{Q\left(\omega^n\right)} dQ(\omega).$$

Now let $L_n = \{\omega : P\left(\omega^n\right) \leq kQ\left(\omega^n\right)\}$. Then

$$\begin{aligned} P\left(E_n \cap L_n \cap A_n^+\right) &= \int_{E_n \cap L_n \cap A_n^+} \frac{P\left(\omega^n\right)}{Q\left(\omega^n\right)} dQ(\omega) \\ &\leq k \int_{E_n \cap L_n \cap A_n^+} dQ \\ &\leq kQ(E_n) < k\varepsilon. \end{aligned}$$

We seek a lower bound for

$$P\left(E_n \cap L_n^c\right) = P\left(E_n\right) - P\left(E_n \cap L_n\right).$$

For each $\omega \in L_n$, either $P\left(\omega^n\right) = 0$ or both $P\left(\omega^n\right) > 0$ and $Q\left(\omega^n\right) > 0$ hold (since $k > 0$). Because $P\left(\left\{\omega : P\left(\omega^n\right) > 0\right\}\right) = 1$, it follows that

$$P\left(L_n\right) = P\left(L_n \cap A_n^+\right).$$

Therefore

$$P\left(E_n \cap L_n\right) = P\left(E_n \cap L_n \cap A_n^+\right).$$

As shown above, $P\left(E_n \cap L_n \cap A_n^+\right) < k\varepsilon$. So, $P\left(E_n \cap L_n\right) < k\varepsilon$. Thus,

$$\begin{aligned} P\left(E_n \cap L_n^c\right) &= P\left(E_n\right) - P\left(E_n \cap L_n\right) \\ &> 1 - \varepsilon - P\left(E_n \cap L_n\right) \\ &> 1 - \varepsilon - k\varepsilon. \end{aligned}$$

Hence,

$$P\left(\left\{\omega : P\left(\omega^n\right) > kQ\left(\omega^n\right)\right\}\right) = P\left(L_n^c\right) \geq P\left(L_n^c \cap E_n\right) > 1 - \varepsilon - k\varepsilon.$$

■

Proof of Theorem 4. Fix $\delta < \frac{\varepsilon^2}{1+\varepsilon}$ and let μ satisfy $\|Q_\mu - P\| > 1 - \delta$ for every $P \in \Lambda$. For every $P \in \Lambda$ let $d_P \in \mathbb{N}$ be such that $\max_{E \in \mathcal{F}_{d_P}} \left|P\left(E\right) - Q_\mu\left(E\right)\right| > 1 - \delta$. Hence, by Lemma 3, for every $d \geq d_P$

$$P\left(\left\{\omega : P\left(\omega^d\right) > \frac{1}{\varepsilon} Q\left(\omega^d\right)\right\}\right) > 1 - \frac{\delta}{\varepsilon} - \delta.$$

By the choice of δ, we have that $\frac{\delta}{\varepsilon} + \delta < \varepsilon$. Hence, a forecaster who chooses a deadline $d \geq d_P$ and then predicts according to P will pass the test with probability, under P, greater than $1 - \varepsilon$.

In order to prove the second part of the theorem we first need to show that T_μ is measurable. As in the proof of Theorem 2, it is sufficient to verify that the map

$(d,P) \mapsto T_{\mu,\varepsilon}(d,\omega,P)$ is measurable for every ω. For every ω, we have

$$\{(d,P) : T_{\mu,\varepsilon}(d,\omega,P) = 1\} = \bigcup_{d\in\mathbb{N}} \left(\{d\} \times \left\{ P : P\left(\omega^d\right) > \frac{1}{\varepsilon} Q_\mu\left(\omega^d\right) \right\} \right)$$

the result then follows from the measurability of the map $P \mapsto P(\omega^n) - \frac{1}{\varepsilon} Q(\omega^n)$.

Now consider a strategic forecaster. For every d and every $P \in \Delta(\Omega)$, Lemma 2 implies

$$Q_\mu(\{\omega : T_{\mu,\varepsilon}(d,\omega,P) = 1\}) = Q_\mu\left(\left\{\omega : \varepsilon P\left(\omega^d\right) > Q_\mu\left(\omega^d\right)\right\}\right) < \varepsilon.$$

That is, for every d and P a forecaster who chooses a deadline d and then predicts according to P passes the test with probability, under Q_μ, lower than ε. It then follows that for every mixed strategy ζ we have

$$\begin{aligned} & (Q_\mu \otimes \zeta)(\{(\omega,(d,P)) : T_{\mu,\varepsilon}(d,\omega,P) = 1\}) \\ = & \int_{\mathbb{N}\times\Delta(\Omega)} Q_\mu(\{\omega : T_{\mu,\varepsilon}(d,\omega,P) = 1\})\, d\zeta(d,P) < \varepsilon. \end{aligned}$$

So, by Fubini's theorem,

$$\int_\Omega \zeta(\{(d,P) : T_{\mu,\varepsilon}(d,\omega,P) = 1\})\, dQ_\mu(\omega) < \varepsilon.$$

Fix a strategy ζ. Let π be the marginal of ζ with respect to $\mathbb{N}$. For each d with $\pi(d) > 0$ let ζ_d be the marginal with respect to $\Delta(\Omega)$ of the conditional probability measure $\zeta(\cdot|\{d\} \times \Delta(\Omega))$. For each d, the function $\omega \mapsto \zeta_d(\{P : T_{\mu,\varepsilon}(d,\omega,P)\} = 1)$ is $\mathcal{F}_d$-measurable. Hence it is continuous. It then follows that

$$\zeta(\{(d,P) : T_{\mu,\varepsilon}(d,\omega,P) = 1\}) = \sum_{d:\pi(d)>0} \pi(d)\, \zeta_d(\{P : T_{\mu,\varepsilon}(d,\omega,P) = 1\})$$

is a continuous function of ω. Therefore, $\int_\Omega \zeta(\{(d,P) : T_{\mu,\varepsilon}(d,\omega,P) = 1\})\, dQ_\mu(\omega)$ is the integral with respect to Q_μ of a continuous function. Hence, it is a continuous function of μ. Because $\mu\left(\overline{\Lambda}\right) = 1$, then the same argument used in the proof of Theorem 2 shows that μ can be approximated by a prior with finite support $\mu_\zeta \in \Delta_o(\Lambda)$ such that

$$\int_\Omega \zeta(\{(d,P) : T_{\mu,\varepsilon}(d,\omega,P) = 1\})\, dQ_{\mu_\zeta}(\omega) < \varepsilon.$$

Thus, there must be a law $P_\zeta \in \Lambda$ in the support of μ_ζ such that

$$\begin{aligned} & (P_\zeta \otimes \zeta)\left(\{(\omega,(d,P)) : T_{\mu,\varepsilon}(d,\omega,P)=1\}\right) \\ = & \int_\Omega \zeta\left(\{(d,P) : T_{\mu,\varepsilon}(d,\omega,P)=1\}\right) dP_\zeta(\omega) < \varepsilon. \end{aligned}$$

■

A.5 Results on Maximal and identifable paradigms

Proof of Remark 1. Let $\{P_1, P_2, ...\}$ be a countable subset of distinct elements of Λ. Given $\varepsilon > 0$, consider a prior $\mu \in \Delta(\Lambda)$ assigning probability 1 to $\{P_1, P_2, ...\}$ and such that $\mu(\{P_i\}) \leq \varepsilon$ for all $1 \leq i < \infty$. For any $P \in \Lambda$, we have $P(\{\omega : \varphi(\omega) = P\}) = 1$ and

$$Q_\mu(\{\omega : \varphi(\omega) = P\}) = \sum_i \mu(P_i) P_i(\{\omega : \varphi(\omega) = P\}) \leq \varepsilon.$$

Thus, $\|P - Q_\mu\| \geq 1 - \varepsilon$. By Theorem 1, Λ is testable. An example of a testable but non-identifiable paradigm is provided in the main text. ■

Given any two sets Y and Z and $D \subseteq Y \times Z$, for every $y \in Y$ and $z \in Z$ we will denote by $D_y = \{z : (y,z) \in D\}$ and $D^z = \{y : (y,z) \in D\}$ the corresponding sections. Given a complete and separable metric space Y, let $\mathcal{B}(Y)$ denote the corresponding Borel sigma-algebra. Finally, recall that two measures P and Q are *orthogonal* if $\|P - Q\| = 1$. Equivalently, if and only if there exists an event E such that $P(E) = 1$ and $Q(E) = 0$. Given a complete and separable metric space Z, let $\mathcal{B}(Z)$ be the sigma-algebra of Borel subsets of Z and $\Delta(Z)$ the space of Borel probability measures on Z.

Theorem 13 (Burgess and Mauldin (1981)) *Let Y and Z be complete and separable metric spaces. Let $m : Y \times \mathcal{B}(Z) \to [0,1]$ be such that $m(y,\cdot)$ belongs to $\Delta(Z)$ for every $y \in Y$ and $m(\cdot,E)$ is Borel for every $E \in \mathcal{B}(Z)$. Then either*

a *There exists a non-empty compact perfect $K \subseteq Y$ and a Borel $D \subseteq Y \times Z$ such that the sections $\{D_y : y \in K\}$ are pairwise disjoint and satisfy $m(y, D_y) = 1$ for every $y \in K$; or*

b *If S is a subset of Y such that the measures $\{m(y,\cdot) : y \in S\}$ are pairwise orthogonal, then S is countable.*

Proof of Theorem 5. Let Λ be a paradigm. By assumption, it is measurable. By Theorem 13.1 in Kechris (1995), there exists a topology τ on $\Delta(\Omega)$ such that Λ is a complete and separable metric space and the Borel sigma-algebra generated by τ is equal to the original Borel sigma-algebra generated by the weak* topology. From now on endow $\Delta(\Omega)$ with the topology τ. We apply the Burgess-Mauldin Theorem, by letting $Y = \Lambda$, $Z = \Omega$ and $m(P, E) = P(E)$ for all $(P, E) \in \Lambda \times B$. We now show that condition (b) in the theorem is violated. Let $\mu \in \Delta(\Lambda)$ be a prior such that $Q_\mu \perp P$ for every $P \in \Lambda$. Let $\boldsymbol{\omega}_1$ be the first uncountable ordinal. We now construct a transfinite sequence $\{P_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} < \boldsymbol{\omega}_1\}$ of measures belonging to Λ and mutually orthogonal. The proof proceeds by induction. Suppose that a transfinite sequence $\{P_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} < \boldsymbol{\beta}\}$ of pairwise orthogonal measures has been defined for some $\boldsymbol{\beta} < \boldsymbol{\omega}_1$. For every $\boldsymbol{\alpha} < \boldsymbol{\beta}$ there exists an event $E_{\boldsymbol{\alpha}} \in \mathcal{B}$ such that $Q_\mu(E_{\boldsymbol{\alpha}}) = 1$ and $P_{\boldsymbol{\alpha}}(E_{\boldsymbol{\alpha}}) = 0$. Let $E_{\boldsymbol{\beta}} = \bigcap_{\boldsymbol{\alpha}<\boldsymbol{\beta}} E_{\boldsymbol{\alpha}}$. By definition, $P_{\boldsymbol{\alpha}}(E_{\boldsymbol{\beta}}) = 0$ for every $\boldsymbol{\alpha} < \boldsymbol{\beta}$. Because $\{\boldsymbol{\alpha} : \boldsymbol{\alpha} < \boldsymbol{\beta}\}$ is countable, then $Q_\mu(E_{\boldsymbol{\beta}}) = 1$. Hence $\int_\Lambda P(E_{\boldsymbol{\beta}})\,\mu(dP) = 1$. So, $\mu(\{P : P(E_{\boldsymbol{\beta}}) = 1\}) = 1$. In particular, there exists a law $P_{\boldsymbol{\beta}} \in \Lambda$ such that $P_{\boldsymbol{\beta}}(E_{\boldsymbol{\beta}}) = 1$. Hence, the measures $\{P_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} < \boldsymbol{\beta}\} \cup \{P_{\boldsymbol{\beta}}\}$ are mutually orthogonal. Proceeding by induction, we obtain a collection $\{P_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} < \boldsymbol{\omega}_1\}$ of mutually orthogonal measures. We conclude that (b) is violated.

Therefore, there exists a compact perfect subset $\Gamma \subseteq \Lambda$ and a Borel $D \subseteq \Lambda \times \Omega$ such that for every P and Q belonging to Γ, if $P \neq Q$ then the sections D_P and D_Q are disjoint and satisfy $P(D_P) = Q(D_Q) = 1$. Let Ω_1 be the projection of D on Ω. We now show that Ω_1 is measurable. Notice that for each ω, the section D^ω contains at most one measure (if $P, Q \in D^\omega$ then $\omega \in D_P \cap D_Q$, but $D_P \cap D_Q = \emptyset$ if $P \neq Q$). It follows then by the Lusin-Novikov theorem (Theorem 18.10, Kechris (1995)) that Ω_1 is in fact Borel. Now fix a measure $Q \in \Gamma$ and define $f : \Omega \to \Gamma$ as $f(\omega) = P$ if $(\omega, P) \in D$ and $f(\omega) = Q$ if $\omega \in \Omega_1^c$. The graph of f is $D \cup (\Omega_1^c \times \{Q\})$, a Borel subset of $\Omega \times \Delta(\Omega)$. Hence, f is Borel (Theorem 14.12, Kechris (1995)). For each $P \in \Gamma$ we have $P(\{\omega : f(\omega) = P\}) \geq P(D_P) = 1$. Hence Γ is identifiable. The first part of the proof is concluded by taking $\tilde{\Lambda}$ to be equal to Γ.

Now assume, in addition, that Λ is closed. Then Λ is a complete and separable metric space. So, we can take τ to be the original weak* topology. The set Γ is compact and perfect, so it contains a subset $\tilde{\Lambda} \subseteq \Gamma$ that is homeomorphic to $\{0,1\}^\infty$ (see Corollary 6.5, Kechris (1995)). The space $\tilde{\Lambda}$ is identifiable: let $E = f^{-1}(\tilde{\Lambda})$ and fix a measure $R \in \tilde{\Lambda}$. Then define $g : \Omega \to \tilde{\Lambda}$ as $g(\omega) = f(\omega)$ if $\omega \in E$; and $g(\omega) = R$ if $\omega \in E^c$. Then, for every Borel subset $\Psi \subseteq \tilde{\Lambda}$, we have $g^{-1}(\Psi) = f^{-1}(\Psi) \cap E$ if

$R \notin \Psi$ and $g^{-1}(\Psi) = (f^{-1}(\Psi) \cap E) \cup E^c$ if $R \in \Psi$. So, g is Borel. In addition, for each $P \in \tilde{\Lambda}$ we have $P(\{\omega : g(\omega) = P\}) = 1$. It is standard to verify that $\{0,1\}^\infty$ is homemorphic to Ω. It is also immediate to prove that the map $\omega \mapsto \delta_\omega$ mapping each path to the corresponding degenerate distribution is an homeomorphism in the weak* topology. Thus, $\tilde{\Lambda}$ is homemorphic to to the class of deterministic distributions. ■

Proof of Theorem 6. As shown by Theorem 2, in order to prove that Λ_P^ε is ε-testable it is enough to find a prior $\mu \in \Delta\left(\overline{\Lambda_P^\varepsilon}\right)$ such that $P = Q_\mu$. Consider the set $N = \{\omega : P(\{\omega\}) = 0\}$. Each $\omega \in N$ satisfies $\delta_\omega \in \Lambda_P^\varepsilon$. Notice that P can have at most countably many atoms, so N is dense. The function $\omega \mapsto \delta_\omega$, $\omega \in \Omega$, is continuous, and so $\{\delta_\omega : \omega \in N\}$ is dense in $\{\delta_\omega : \omega \in \Omega\}$. We can therefore conclude that $\{\delta_\omega : \omega \in \Omega\} \subseteq \overline{\Lambda_P^\varepsilon}$. Consider now the prior defined as $\mu(\Gamma) = P(\{\omega : \delta_\omega \in \Gamma\})$ for every measurable set $\Gamma \subseteq \Delta(\Omega)$. Standard arguments shows that μ is well defined and satisfies $Q_\mu = P$. Because $\mu(\{\omega : \delta_\omega \in \Omega\}) = 1$, then $\mu \in \Delta\left(\overline{\Lambda_P^\varepsilon}\right)$. Therefore, Λ_P^ε is ε-testable.

Suppose, as a way of contradiction, that $\Lambda_P^\varepsilon \subseteq \Lambda$, where Λ is a paradigm that is ε'-testable and $\varepsilon' < \frac{\varepsilon}{2}$. As shown in the proof of Theorem 1, there exists a prior $\nu \in \Delta\left(\overline{\Lambda}\right)$ such that $\|Q_\nu - Q\| \geq 1 - 2\varepsilon'$ for every $Q \in \Lambda$. Equivalently,

$$\{Q \in \Delta(\Omega) : \|Q - Q_\nu\| < 1 - 2\varepsilon'\} \subseteq \Lambda^c$$

By assumption, $\Lambda^c \subseteq (\Lambda_P^\varepsilon)^c = \{Q \in \Delta(\Omega) : \|Q - P\| \leq 1 - \varepsilon\}$, so

$$\{Q \in \Delta(\Omega) : \|Q - Q_\nu\| < 1 - 2\varepsilon'\} \subseteq \{Q \in \Delta(\Omega) : \|Q - P\| \leq 1 - \varepsilon\}. \tag{17}$$

To show that this leads to a contradiction, let $R \in \Delta(\Omega)$ be a measure such that $\|R - Q_\nu\| = \|R - P\| = 1$. For instance, let $R = \delta_\omega$ for some path ω that is not an atom of either Q_ν or P. Fix $t \in (2\varepsilon', \varepsilon)$ and consider the measure $tQ_\nu + (1-t)R$. We have

$$\|tQ_\nu + (1-t)R - Q_\nu\| = (1-t)\|R - Q_\nu\| = (1-t) < 1 - 2\varepsilon'.$$

Hence, it follows from (17) that $\|tQ_\nu + (1-t)R - P\| \leq 1 - \varepsilon$. Now let E be an event such that $R(E) = 1$ and $Q_\nu(E) = P(E) = 0$. Then

$$1 - \varepsilon \geq \|tQ_\nu + (1-t)R - P\| \geq tQ_\nu(E) + (1-t)R(E) - P(E) = 1 - t.$$

By construction, $1 - t > 1 - \varepsilon$. So we obtain a contradiction. ■

Proof of 7. By Theorem 1, there exists a prior $\mu \in \Delta\left(\overline{\Lambda}\right)$ such that $\Lambda \subseteq \Lambda^{\varepsilon}_{Q_\mu}$. Using the fact that $\|Q_\mu - P\| > 1-\varepsilon$ for every $P \in \Lambda^{\varepsilon}_{Q_\mu}$, as in the proof of Theorem 2 we obtain a likelihood-ratio test T such that $P\left(\{\omega : T(\omega, P) = 1\}\right) > 1-\varepsilon$ and $Q_\mu\left(\{\omega : T(\omega, P) = 1\}\right) < \varepsilon$ for every $P \in \Lambda^{\varepsilon}_{Q_\mu}$. The proof that for every ζ there exists a law $P_\zeta \in \Lambda$ such that $E_{P_\zeta \otimes \zeta}[T] \leq \varepsilon$ follows then by replicating the argument in the proof of Theorem 2. ■

A.6 Other Proofs

Proof of Theorem 8. Let $N_n(\omega)$ be the number of periods the outcome x has occurred along the path ω up to time n. Finally, given a transition probability π, denote by $E_{x,\pi}[\tau]$ the expected time it will take a Markov chain with transition π and starting from x to return to outcome x. It is a standard fact that for every Markov chain $P_{\rho,\pi}$,

$$\lim_{n\to\infty} \frac{1}{n} N_n(\omega) = \frac{1}{E_{x,\pi}[\tau]} \quad \text{for } P_{\rho,\pi}\text{-almost all paths } \omega \text{ such that } \tau(\omega) < \infty.$$

(see Proposition 6.2.49 in Dembo, 2016). That is, along a path ω where x occurs at least once, the frequency $\frac{1}{n}N_n(\omega)$ converges to the constant $c(\pi) = E_{x,\pi}[\tau]^{-1}$. For each $k \in \mathbb{R}$, let E_k to be the set of paths ω along which the fraction $\frac{1}{n}N_n(\omega)$ converges to k. So, each Markov law $P_{\rho,\pi}$ satisfies $P_{\rho,\pi}\left(E_0 \cup E_{c(\pi)}\right) = 1$. We now return to the prior μ. For each α, the constant $c(\pi_\alpha)$ can be easily computed to be $2-\alpha$. The key fact is that $c(\pi_\alpha) \neq c(\pi_{\alpha'})$ whenever $\alpha \neq \alpha'$. Therefore, given any Markov $P_{\rho,\pi}$ there is at most one $\alpha \in [0,1]$ such that $c(\pi_\alpha) = c(\pi)$. Because $P_\alpha\left(E_{c(\pi_\alpha)}\right) = 1$ for every α then

$$Q_\mu\left(E_0 \cup E_{c(\pi)}\right) = \int_0^1 P_\alpha\left(E_0 \cup E_{c(\pi)}\right) d\alpha = 0.$$

Because $P_{\rho,\pi}\left(E_0 \cup E_{c(\pi)}\right) = 1$ then $\|Q_\mu - P_{\pi,\rho}\| = 1$. So, by Theorem 1, the prior μ guarantees that Λ is testable. ■

Proof of Theorem 9. Let Λ be ε-testable in n periods. Then, by substituting the total-variation distance with the semi-distance $\|Q-P\|_n$ and applying the same arguments used in the proof of Theorem 1 it follows that there exists a prior $\mu \in \Delta\left(\overline{\Lambda}\right)$ such that $\|Q_\mu - P\|_n > 1-2\varepsilon$ for all $P \in \Lambda$. Only one change is necessary: the same results in Shiryaev (2016) cited in the proof of Theorem 1 imply $\|P-Q\|_n =$

$\max_\phi \left| \int_\Omega \phi dP - \int_\Omega \phi dQ \right|$ where the maximum is taken over all functions $\phi : \Omega \to [0,1]$ that are $\mathcal{F}_n$-measurable.

Conversely, let $\mu \in \Delta\left(\overline{\Lambda}\right)$ be a prior such that $\|Q_\mu - P\|_n > 1 - \varepsilon$ for all $P \in \Lambda$. The first part of the proof follows, verbatim, the proof of Theorem 2 (notice that by assumption $n_P \leq n$ for every $P \in \Lambda$). ■

Proof of Theorem 10. Let μ be a prior such that $\|Q_\mu - P\|_n > 1 - \delta$ for every $P \in \Lambda$. Then, Lemma 3 implies

$$P\left(\left\{\omega : P\left(\omega^n\right) > \frac{1}{\varepsilon} Q_\mu\left(\omega^n\right)\right\}\right) > 1 - \frac{\delta}{\varepsilon} - \delta > 1 - \varepsilon$$

Hence, the test does not reject the truth with probability $1 - \varepsilon$. By Lemma 2, $Q_\mu\left(\left\{\omega : P\left(\omega^n\right) > \frac{1}{\varepsilon} Q_\mu\left(\omega^n\right)\right\}\right) < \varepsilon$. That the test is ε-nonmanipulable follows by replicating, with only notational changes, the proof of Theorem 4. ■

B Appendix: Characterization of uniform testability

The purpose of this section is to provide necessary and sufficient conditions for a paradigm to be uniformly testable in the sense of Definition 10.

Definition 11 *Given $\delta > 0$ and a paradigm Λ of $\Delta(\Omega)$, a sequence of priors (μ_n) in $\Delta\left(\overline{\Lambda}\right)$ is* δ-separating *if:*

1. *there exists a sequence (ε_n) such that $\varepsilon_n \downarrow 0$ and $\left\|Q_{\mu_n} - P\right\| \geq 1 - \varepsilon_n$ for every $P \in \Lambda$; and*

2. *there exists a constant $\lambda > 0$ such that $\inf_n \mu_n\left(B_\delta(P)\right) \geq \lambda$ for every $P \in \Lambda$.*

Property (1) mirrors the characterization of Theorem 1. Property (2) imposes a uniform bound, along the sequence, on the probability of each ball of radius δ. The next theorem shows that this property characterizes uniform testability.

Theorem 14 *Let Λ be a paradigm. If there exists a sequence of priors (μ_n) that is δ-separating then Λ is uniformly testable with precision δ. Conversely, if Λ is uniformly testable with precision δ then there exist a sequence of priors (μ_n) that is δ'-separating for every $\delta' > \delta$.*

Given this result, Theorem 11 follows from the following lemma.

Lemma 4 *Let $\mu \in \Delta(\Delta(\Omega))$ be a prior and let $\Gamma \subseteq \Delta(\Omega)$ be its support. For every $\delta > 0$ there exists a constant $\lambda > 0$ such that*

$$\mu(B_\delta(P)) \geq \lambda \text{ for all } P \in \Gamma.$$

The result implies Theorem 11: Given a prior μ with support $\overline{\Lambda}$ and such that $\|Q_\mu - P\| = 1$ for every $P \in \Lambda$, the constant sequence (μ_n) where $\mu_n = \mu$ for every n is δ-separating for every $\delta > 0$. Hence, by Theorem 14 the paradigm Λ is uniformly testable with precision δ for every $\delta > 0$.

B.1 Proofs

Proof of Lemma 4. Suppose not. Then there must exist a sequence (P_n) in Γ such that $\mu(B_\delta(P_n)) \to 0$ as $n \to \infty$. Using the compactness of Γ we can assume (taking a subsequence if necessary) that P_n converges to a law $P \in \Gamma$. Denote by d the distance fixed to metrize the weak* topology. For each law Q, if $Q \in B_\delta(P)$ then $d(P_n, Q) < \delta$ for all n large enough. Thus $Q \in B_\delta(P_n)$ for all n large enough. Thus,

$$1_{B_\delta(P)}(Q) \leq \liminf_{n\to\infty} 1_{B_\delta(P_n)}(Q) \text{ for every } Q \in \Gamma$$

where $1_{B_\delta(P)}$ denotes the indicator function of $B_\delta(P)$. By applying Fatou's lemma, we can then conclude that

$$\mu(B_\delta(P)) \leq \int_{\Delta(\Omega)} \liminf_{n\to\infty} 1_{B_\delta(P_n)} d\mu \leq \liminf_n \mu(B_\delta(P_n)) = 0$$

Hence $\mu(B_\delta(P)) = 0$. Since $P \in \Gamma$, then μ must satisfy $\mu(B_\gamma(P)) > 0$ for every $\gamma > 0$ so we reach a contradiction, and the proof is finished. ■

Proof of Theorem 14. Let (μ_n) be a sequence of priors that is δ-separating. Let $\lambda > 0$ be such that $\inf_n \mu_n(B_\delta(P)) \geq \lambda$ for every $P \in \Lambda$. As shown in the proof of Theorem 2, we can find for every n a finite test T_n with the properties that T_n does not reject the truth with probability $1 - \varepsilon_n$ and for every strategy ζ,

$$\int_{\overline{\Lambda}} E_{P\otimes\zeta}[T_n] \, d\mu_n(P) \leq \varepsilon_n.$$

By applying Markov's inequality we obtain

$$\mu_n\left(\left\{P \in \overline{\Lambda} : E_{P\otimes\zeta}\left[T_n\right] \leq k\varepsilon_n\right\}\right) \geq 1 - \frac{1}{k\varepsilon_n} \geq 1 - \frac{1}{k} \text{ for all } \zeta.$$

Fix $\varepsilon > 0$ and choose k large enough such that both $\frac{1}{k} \leq \varepsilon$ and $1 - \frac{1}{k} + \lambda > 1$ hold. In addition, given k choose N large enough such that $k\varepsilon_n \leq \varepsilon$ for all $n > N$. Now fix a particular $n > N$. Given $P_o \in \Lambda$ and a strategy ζ, we have

$$\begin{aligned}
& \mu_n\left(\left\{P \in \overline{\Lambda} : E_{P\otimes\zeta}\left[T_n\right] \leq \varepsilon\right\} \cap B_\delta\left(P_o\right)\right) \\
\geq\ & \mu_n\left(\left\{P \in \overline{\Lambda} : E_{P\otimes\zeta}\left[T_n\right] \leq k\varepsilon_n\right\} \cap B_\delta\left(P_o\right)\right) \\
=\ & \mu_n\left(\left\{P \in \overline{\Lambda} : E_{P\otimes\zeta}\left[T_n\right] \leq k\varepsilon_n\right\}\right) \\
& + \mu_n\left(B_\delta\left(P_o\right)\right) - \mu_n\left(\left\{P \in \overline{\Lambda} : E_{P\otimes\zeta}\left[T_n\right] \leq k\varepsilon_n\right\} \cup B_\delta\left(P_o\right)\right) \\
\geq\ & 1 - \frac{1}{k} + \lambda - 1 > 0.
\end{aligned}$$

This implies we can select a measure $Q_\zeta \in \overline{\Lambda} \cap B_\delta\left(P_o\right)$ such that $E_{Q_\zeta\otimes\zeta}\left[T_n\right] \leq \varepsilon$. By continuity of the map $P \mapsto E_{P\otimes\zeta}\left[T_n\right]$ we can then select a measure $P_\zeta \in \Lambda \cap B_\delta\left(P_0\right)$ such that $E_{P\otimes\zeta}\left[T_n\right]$. Because P_o is arbitrary, then it follows that the test T_n satisfies the conditions of Definition 10. Because ε is arbitrary, it follows that Λ is uniformly testable with precision δ.

Now, let Λ be uniformly testable with precision δ. Fix $\varepsilon > 0$. Then we can find a test T such that

$$\sup_{\zeta\in\Delta(\Delta(\Omega))} \inf_{Q\in\Lambda\cap B_\delta(P_o)} E_{Q\otimes\zeta}\left[T\right] \leq \varepsilon$$

for every $P_o \in \Lambda$. By replicating the proof of Theorem 1 we obtain for each $P_o \in \Lambda$ and every $\varepsilon > 0$ a prior $\mu\left[\varepsilon, P_o\right] \in \Delta\left(\overline{\Lambda \cap B_\delta\left(P_o\right)}\right)$ such that $\left\|Q_{\mu[\varepsilon,P_o]} - P\right\| \geq 1 - 2\varepsilon$ for every $P \in \Lambda$. Let $\{P^1, P^2, ...\}$ be a countable dense subset of Λ. Define now the priors

$$\pi = \sum_i 2^{-i}\delta_{P^i} \text{ and } \mu\left[\varepsilon\right] = \sum_i 2^{-i}\mu\left[\varepsilon, P^i\right].$$

Fix $\delta' > \delta$ and $P_o \in \Lambda$. We have

$$\mu\left[\varepsilon\right]\left(B_{\delta'}\left(P_o\right)\right) = \sum_i 2^{-i}\mu\left[\varepsilon, P_i\right]\left(B_{\delta'}\left(P_o\right)\right)$$

By assumption $\mu[\varepsilon, P_i]\left(\overline{B_\delta(P_i)}\right) = 1$ for every i. Hence,

$$\begin{aligned}\mu[\varepsilon](B_{\delta'}(P_o)) &\geq \mu[\varepsilon]\left(\left\{P_i : \overline{B_\delta(P_i)} \subseteq B_{\delta'}(P_o)\right\}\right) \\ &= \pi\left(\left\{P_i : \overline{B_\delta(P_i)} \subseteq B_{\delta'}(P_o)\right\}\right).\end{aligned}$$

It follows from the triangle inequality that $B_{\delta'-\delta}(P_o) \subseteq \left\{Q \in \Delta(\Omega) : \overline{B_\delta(Q)} \subseteq B_{\delta'}(P_o)\right\}$. Therefore $\mu[\varepsilon](B_{\delta'}(P_o)) \geq \pi(B_{\delta'-\delta}(P_o))$. Now we can apply Lemma 4 to π and conclude that there exists a constant $\lambda_{\delta'-\delta} > 0$, independent of P_o and ε, such that

$$\mu[\varepsilon](B_{\delta'}(P_o)) \geq \pi(B_{\delta'-\delta}(P_o)) \geq \lambda_{\delta'-\delta}.$$

The last step consists in estimating the distance $\left\|Q_{\mu[\varepsilon]} - P\right\|$ given $P \in \Lambda$. For each i we have $\left\|Q_{\mu[\varepsilon,P_i]} - P\right\| \geq 1-2\varepsilon$ so we can find an event $E_i \in \mathcal{B}$ such that $Q_{\mu[\varepsilon,P_i]}(E_i) \geq 1-2\varepsilon$ and $P(E_i) \leq 2\varepsilon$. For every $n \in \mathbb{N}$, the distance $\left\|Q_{\mu[\varepsilon]} - P\right\| = \max_{E\in\mathcal{B}} Q_{\mu[\varepsilon]}(E) - P(E)$ satisfies

$$\begin{aligned}\left\|Q_{\mu[\varepsilon]} - P\right\| &= \max_{E\in\mathcal{B}} \sum_{i=1}^{\infty} 2^{-i} Q_{\mu[\varepsilon,P_i]}(E) - P(E) \\ &\geq \max_{E\in\mathcal{B}} \sum_{i=1}^{n} 2^{-i} Q_{\mu[\varepsilon,P_i]}(E) - P(E) \\ &\geq \sum_{i=1}^{n} 2^{-i} Q_{\mu[\varepsilon,P_i]}\left(\bigcup_{l=1}^{n} E_l\right) - P\left(\bigcup_{l=1}^{n} E_l\right) \\ &\geq \left(1-2^{-n}\right)(1-2\varepsilon) - 2n\varepsilon\end{aligned} \tag{18}$$

where the last inequality follows from

$$Q_{\mu[\varepsilon,P_i]}\left(\bigcup_{l=1}^{n} E_l\right) \geq Q_{\mu[\varepsilon,P_i]}(E_i) \geq 1-2\varepsilon \text{ and } P\left(\bigcup_{l=1}^{n} E_l\right) \leq \sum_{l=1}^{n} P(E_l) \leq n2\varepsilon.$$

Notice that the lower bound $(1-2^{-n})(1-2\varepsilon) - 2n\varepsilon$ does not depend on P.

Now let $\varepsilon_n = 2^{-n}$ for each n, and consider the sequence of priors $(\mu[\varepsilon_n])$. As shown above, they satisfy $\mu[\varepsilon_n](B_{\delta'}(P_o)) \geq \lambda_{\delta'-\delta} > 0$ for every n and every $P_o \in \Lambda$. Moreover, substituting $\varepsilon = 2^{-n}$ in (18) we obtain $\left\|Q_{\mu[\varepsilon_n]} - P\right\| \geq \left(1-2^{-n+1}\right)^2 - n2^{-n+1}$ for all $P \in \Lambda$. So, $\inf_{P\in\Lambda} \left\|Q_{\mu[\varepsilon_n]} - P\right\| \to 1$ as $n \to \infty$. This concludes the proof that the sequence $(\mu[\varepsilon_n])$ is δ'-separating for every $\delta' > \delta$. ■

References

Aliprantis, C. D., and K. Border. (2006). *Infinite dimensional analysis: a hitchhiker's guide.* Springer.

Alkema, L., Raftery, A. E., & Clark, S. J. (2007). "Probabilistic projections of HIV prevalence using Bayesian melding." *The Annals of Applied Statistics*, 229-248.

Al-Najjar, N.I., Sandroni, A., Smorodinsky, R. and J. Weinstein (2010). "Testing theories with learnable and predictive representations." *Journal of Economic Theory*, 145(6), 2203-2217.

Al-Najjar, N., and E. Shmaya (2016). "Learning the ergodic decomposition." *mimeo.*

Al-Najjar, N., and J. Weinstein (2008). "Comparative testing of experts." *Econometrica*, 76(3), 541-559.

Babaioff, M., L. Blumrosen, N. Lambert and O. Reingold (2011). "Only valuable experts can be valued." *Proceedings of the 12th ACM conference on electronic commerce*, 221-222.

Bergemann, D., and K. Schlag (2011). "Robust monopoly pricing." *Journal of Economic Theory,* 146, 2527–2543.

Blackwell, D. (1980). "There are no Borel SPLIFs." *The Annals of Probability*, 8(6), 1189-1190.

Burgess, J. P., and R.D. Mauldin (1981). "Conditional distributions and orthogonal measures." *The Annals of Probability*, 9(5), 902-906.

Carvajal, A. (2009) "Statistical calibration: A simplification of Foster's proof," *Mathematical Social Sciences*, 58(2), 272-277.

Cerreia-Vioglio, S., Maccheroni, F. and M. Marinacci (2013). "Classical subjective expected utility." *Proceedings of the National Academy of Sciences,* 110(17), 6754-6759.

Corradi, V., and N. R. Swanson (2006). "Predictive density evaluation." *Handbook of economic forecasting*, 1, 197-284.

Dawid, A. P. (1982). "The well-calibrated Bayesian." *Journal of the American Statistical Association*, 77(379), 605-610.

Dawid, A. P. (1984). "Statistical theory: The prequential approach." *Journal of the Royal Statistical Society A*, 147.

Dekel, E. and Y. Feinberg (2006). "Non-Bayesian testing of a stochastic prediction." *Review of Economic Studies*, **73**, 893 - 906.

Diebold, F. X., Tay, A. S., and K. F. Wallis. (1997). "Evaluating density forecasts of inflation: the Survey of Professional Forecasters." *National bureau of economic research.*

Doob, J. L. (1949). "Application of the theory of martingales." *Le calcul des probabilites et ses applications*, 23-27.

Dynkin, E. B. (1978). "Sufficient statistics and extreme points." *The Annals of Probability*, 705-730

Echenique, F. and E. Shmaya (2007). "You won't harm me if you fool me." *mimeo.*

Fan, K. (1953). "Minimax theorems." *Proceedings of the National Academy of Sciences*, 39(1), 42-47.

Feinberg, Y., and C. Stewart (2008). "Testing multiple forecasters." *Econometrica*, 76(3), 561-582.

Feinberg, Y., and N. Lambert (2015). "Mostly calibrated." *International Journal of Game Theory,* 44:153-163.

Fortnow, L., and R. Vohra (2009). "The complexity of forecast testing." *Econometrica*, 77(1), 93-105.

Foster, D. P. (1999). "A proof of calibration via Blackwell's approachability theorem," *Games and Economic Behavior*, 29, 7378.

Foster, D. P., and R. Vohra (1998). "Asymptotic calibration." *Biometrika,* 85(2), 379-390.

Foster, D. P., and R. Vohra (2011). "Calibration: Respice, adspice, prospice." *mimeo.*

Fudenberg, D. and D. Levine (1999). "An easier way to calibrate," *Games and Economic Behavior*, 29, 131-137.

Gilboa, I. and D. Schmeidler (1989). "Maxmin expected utility with non-unique prior." *Journal of mathematical economics*, 18(2), 141-153.

Gneiting, T., and A. E. Raftery (2005). "Weather forecasting with ensemble methods." *Science*, 310 (5746), 248-249.

Gradwohl, R. and Y. Salant (2011). "How to buy advice." *mimeo.*

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

Hewitt, E., and Savage, L. J. (1955). "Symmetric measures on Cartesian products," *Transactions of the American Mathematical Society*, 80(2), 470-501.

Hu, T., and E. Shmaya (2013). "Expressible inspections." *Theoretical Economics*, 8(2) 263-280.

Huber, P. J. (1981). *Robust statistics*. Wiley.

Jackson, M., E. Kalai and R. Smorodinsky (1999). "Bayesian representation of stochastic processes under learning: de Finetti revisited." *Econometrica*, 67(4), 875-893.

Jordan T.H., Chen Y.T., Gasparini P., Madariaga R., Main I., et al. (2011). "Operational earthquake forecasting: state of knowledge and guidelines for utilization." *Annals of Geophysics* 54, 315–91.

Kalai, E., E. Lehrer and R. Smorodinsky. (1999). "Calibrated Forecasting and Merging," *Games and Economic Behavior*, 29(1-2), 151-169.

Kechris, A. (1995). *Classical descriptive set theory*. Springer.

Lehmann, E. L., and J.P. Romano (2006). *Testing statistical hypotheses.* Springer.

Lehrer, E. (2001). "Any inspection is manipulable." *Econometrica*, 69(5), 1333-1347.

Mannor, S. and G. Stoltz (2009) "A geometric proof of calibration," *Mathematics of Operations Research*, 35.4, 721-727.

Neyman, J. and E. S. Pearson. (1933). "The testing of statistical hypotheses in relation to probabilities a priori." *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 29, No.4.

Olszewski, W. (2015). "Calibration and expert testing." *Handbook of Game Theory with Economic Applications,* 4, 949-984.

Olszewski, W. and M. Peski (2011). "The principal-agent approach to testing experts." *American Economic Journal: Microeconomics* 3, 89-113

Olszewski, W., and A. Sandroni. (2008). "Manipulability of future-independent tests." *Econometrica*, 76(6), 1437-1466.

Olszewski, W., and A. Sandroni. (2009). "Strategic manipulation of empirical tests." *Mathematics of Operations Research*, 34.1, 57-70.

Olszewski, W., and A. Sandroni. (2009b). "A nonmanipulable test." *The Annals of Statistics*. 1013-1039.

Olszewski, W., and A. Sandroni (2011). "Falsifiability." *American Economic Review*, 101, 788-818.

Phelps, R. R. (2001). *Lectures on Choquet's theorem*. Springer.

Raftery A.E., Li N., Sevcikova H., Gerland P., Heilig G.K. (2012). "Bayesian probabilistic population projections for all countries." *Proceedings of the National Academy of Sciences*, 109, 13915–21

Sandroni, A. (2003). "The reproducible properties of correct forecasts." *International Journal of Game Theory*, 32(1), 151-159.

Sandroni, A. and E. Shmaya (2014). "A prequential test for exchangeable theories." *Journal of Dynamics and Games*, 1(3), 497-505.

Sandroni, A., Smorodinsky, R., and Vohra, R. V. (2003). "Calibration with many checking rules." *Mathematics of Operations Research*, 28(1), 141-153.

Shmaya, E. (2008). "Many inspections are manipulable." *Theoretical Economics*, 3(3), 367-382.

Shiryaev, A.N. (1996). *Probability*, Springer.

Starr, R. M. (1969). "Quasi-equilibria in markets with non-convex preferences." *Econometrica*, 37(1), 25-38.

Stewart, C. (2011). "Nonmanipulable bayesian testing." *Journal of Economic Theory.* 146(5), 2029-2041.

Tetlock P.E. (2005). *Political Expert Judgement*. Princeton.

Timmermann, A. (2000). "Density forecasting in economics and finance." *Journal of Forecasting*, 19(4), 231-234.

Weizsäcker, H. V. (1996). "Some reflections on and experiences with SPLIFs." *Lecture Notes-Monograph Series.* 391-399.