Calibration: Respice, Adspice, Prospice

Dean P. Foster* Rakesh V. Vohra†

Abstract

"Those who claim for themselves to judge the truth are bound to possess a criterion of truth."

Sextus Empiricus

1 Introduction

Suppose one is asked to forecast the probability of rain on successive days. How should one assess the accuracy of the forecast? If one forecasts a 25% chance of rain and it rains, was the forecast in error?

A popular criteria for judging the effectiveness of a probability forecast is called *calibration*. Dawid [5] offers the following intuitive definition of calibration:

> "Suppose that, in a long (conceptually infinite) sequence of weather forecasts, we look at all those days for which the forecast probability of precipitation was, say, close to some given value ω and (assuming these form an infinite sequence) determine the long run proportion p of such days on which the forecast event (rain) in fact occurred. The plot of p against ω is termed the forecaster's *empirical calibration curve*. If the curve is the diagonal $p = \omega$, the forecaster may be termed (empirically) *well calibrated*."

Notice, the calibration criterion relies only on the realized forecasts and outcomes to make a determination. It assumes the data will speak for itself.

The calibration criterion is used, for example, to assess the accuracy of prediction markets, see Page and Clemen (2010) [30]. Philip Tetlock [37] uses it in his comprehensive analysis of pundits. We quote from a 2006 blog entry by Tetlock [38]:

> "Between 1985 and 2005, boomsters made 10-year forecasts that exaggerated the chances of big positive changes in both financial markets (e.g., a Dow Jones Industrial Average of 36,000) and world politics (e.g., tranquility in the Middle East and dynamic growth in sub-Saharan Africa). They assigned probabilities of 65% to rosy scenarios that materialized only 15% of the time.

*Department of Statistics, University of Pennyslvannia, The Wharton School, Philadlephia PA 19104.

†Department of Managerial Economics and Decision Sciences, Kellogg Graduate School of Management, Northwestern University, Evanston IL 60208.

> In the same period, doomsters performed even more poorly, exaggerating the chances of negative changes in all the same places where boomsters accentuated the positive, plus several more (I still await the impending disintegration of Canada, Nigeria, India, Indonesia, South Africa, Belgium, and Sudan). They assigned probabilities of 70% to bleak scenarios that materialized only 12% of the time."

Fans of Isaiah Berlin will be interested to know that Tetlock concludes that foxes are better calibrated than hedgehogs.[1]

2 Notation

The intuitive definition of calibration is meaningless when no forecast is ever repeated. One way around this is to base the definition on what is known as the calibration component of the Brier score (see [3] and [23]). To describe it we introduce notation. Let $S = \{0, 1\}$ be the state space.[2] We can think of '1' as recording the state 'rain'. An element of S is called an *outcome*. Let S^n, for $n \in N$, be the n-Cartesian product of S and S^* the set of all infinite 0-1 sequences. An n-sequence of outcomes is denoted $s = (s_1, s_2, \ldots, s_n) \in S^n$ where s_i denotes the state realized in period i. An infinite sequence is denoted s^*. Given $s \in S^n$ and $r < n$, let $s^r = (s_1, s_2, \ldots, s_r) \in S^r$ be the prefix of length r of s.

An element of $[0, 1]$ is called a *forecast* of the event '1.' A forecast made in period r refers to outcomes that will be observed in period $r + 1$. Let Δ^* be the set of probability distributions over $[0, 1]$. A forecasting algorithm is a function:

$$F : \bigcup_{r=0}^{n-1} (S^r \times [0, 1]^r) \to \Delta^*$$

At the end of each period $r < n$, an r-history $(s^r, f_0, f_1, \ldots, f_{r-1}) \in S^r \times [0, 1]^r$ is observed. Here $f_j \in [0, 1]$ is the forecast made by F in period j. Let $f^r = (f_0, \ldots, f_r)$. Based on this r-history, the forecaster must decide which forecast $f_r \in [0, 1]$ to make in period r. The forecaster is allowed to randomize. So, $f_r \in [0, 1]$ can be selected (possibly) at random, using a probability distribution in Δ^*.

Let $n_t(p; F, s^*)$ be the number of times F forecasts p up to (but not including) time t on the sequence s^*. Let $\rho_t(p; F, s^*)$ be the fraction of those times that it actually rained. In other words,

$$\begin{aligned} n_t(p; F, s^*) &\equiv \sum_{r=0}^{t-1} I_{f_r=p}, \\ \rho_t(p; F, s^*) &\equiv \sum_{r=0}^{t-1} \frac{s_{r+1} I_{f_r=p}}{n_t(p; F, s^*)}, \end{aligned}$$

[1] Berlin offered a classification of thinkers inspired by a fragment of poetry due to Archilochus. Rendered in English, it reads: The fox knows many things, but the hedgehog knows one big thing. Foxes are thinkers who draw on a variety of perspectives to understand the world. Hedgehogs, believe that the world can only be understood through a single perspective.

[2] The results extend easily to more than two states.

where I is the indicator function. In the definition it is convenient to assume that F is restricted to selecting forecasts from a finite set, A, fixed a priori. The requirement that F select from a fixed set A is not a severe restriction for practical purposes. Many weather forecasters, for example, forecast probabilities to only one decimal place.

The calibration score of F with respect to s^* after t periods is denoted $C_t(F, s^*)$ where

$$C_t(F, s^*) = \sum_{p \in A} \left(\rho_t(p; F, s^*) - p\right)^2 \frac{n_t(p; F, s^*)}{t}$$

Thus, F is well calibrated with respect to s^* if and only if $C_t(F, s^*)$ goes to zero as t goes to infinity.

3 Calibrated Forecasts

While a forecast that reports the correct probabilities (conditional on the history) in each period, will have a low calibration score, what about an 'incorrect' forecast? Foster and Vohra [10] exhibit a randomized forecasting algorithm that almost surely will be calibrated on all sequences s^*. No assumption is made about the process that generates s^*. The 'almost surely' in the statement refers to the distribution induced by the randomization within the forecasting algorithm. At first blush this result appears surprising, so we outline a proof of existence based on the mini-max theorem.[3]

3.1 Existence

This existence proof constructs a zero-sum game played between the forecaster and 'Nature.' Fix the number of periods to be t. So that the forecaster's strategy space is finite, we restrict him to picking one of the following in each period as a forecast: $0, 1/k, 2/k, \ldots, 1$. Here k is a sufficiently large integer to be chosen later. A pure strategy for the forecaster will consist of a t-vector of forecasts, where each element of the vector will be of the form j/k for $0 \leq j \leq k$.[4] Thus, his strategy space consists of $(k+1)^{2^t-1}$ pure strategies. Nature's strategy space is the set of all 2^t binary sequences.[5] If Nature picks s^t, then the forecaster's 'loss' from a particular sequence of forecasts is the calibration score of that sequence of forecasts with respect to s^t.

Now suppose that Nature picks a (possibly randomized) strategy first. Assume that the forecaster knows the randomization strategy that Nature will follow but not the realization. To use the minimax theorem we need to specify a strategy for the forecaster which will keep his calibration score less than ϵ. If we can do this for all possible mixed strategies of Nature, then, by the mini-max theorem, there must exist a mixed strategy for the forecaster which will guarantee him a calibration score less than ϵ.

Given each mixed strategy of Nature, the forecaster can compute the conditional probability of the next term in the sequence being a '1'. The forecast of the corresponding term will be obtained

[3] There are a host of other proofs. See, for example, [13], [15], [11], [4] and [21].

[4] For economy of exposition only, we assume that a forecast in each period cannot depend on what happened in the past.

[5] This assumes that Nature's strategy in each period cannot depend on what she saw in the previous period. The argument is the same if we drop this restriction.

by rounding this probability to the nearest i/k value. Assuming that k is much less than $n^{1/3}$ his calibration score will be less than $1/k$. Here is an outline of why this must be so. The forecasters calibration score is

$$C_t(F, s^t) = \sum_{j=0}^{k} \Big(\rho_t(j/k; F, s^t) - j/k\Big)^2 \, \frac{n_t(j/k; F, s^t)}{t}.$$

Now look at all the times the forecaster forecast j/k. He did so because the probability that Nature would pick a 1 on that round was some number q with the property that $|q - i/k|$ was minimized for $i = j$. This implies that $|q - j/k| \leq 1/k$. By a law of large numbers[6] argument we would expect that $|\rho_t(j/k; F, s^t) - j/k| \leq 1/k$. Hence

$$C_t(F, s^t) \leq \sum_{j=0}^{k} (1/k)^2 \frac{n_t(j/k; F, s^t)}{t} = 1/k.$$

Thus, there exists a randomized strategy which will guarantee him a calibration score of at most $1/k$. Randomization is essential. While a malevolent nature may be able to make one forecaster look bad according to the calibration criterion, it is harder for it to make many forecasters look bad at the same time. To quote Schervish [34]:

> "The more different forecasts that nature is trying to make look bad, the more flexibility all forecasters have to try to look good."

3.2 Extensions

Under the calibration criterion, a forecaster with no meteorological knowledge would be indistinguishable from one who knew the distribution that governs the change in weather. Is this surprising? In a sense no, since calibration by itself is not a sufficient condition for a forecast to be deemed good. To see this, consider the sequence of outcomes and forecasts below.

outcome	0	1	0	1	0
forecast	0.5	0.5	0.5	0.5	0.5

Assuming the sequence of outcomes and forecasts repeats indefinitely, the forecast will be calibrated with respect to this sequence. However, the forecast displayed is not the only forecast that will be calibrated with respect to this sequence. For example, the forecast $0, 1, 0, 1 \ldots$, is calibrated with respect to the sequence of outcomes displayed. Thus, calibration isn't sufficient to distinguish good from excellent forecasters. Neverthless, one can agree that if someone forecasted ".7" in each period on the above sequence, their poor calibration would be one way to describe it as a bad forecast.

We can demand more by breaking the sequence into two subsequences; one corresponding to even periods and the other to odd periods, and require the forecast to match the frequency on each subsequence. Consider the table below.

outcome	0	0	1	1	0	0	1	1	0	0
forecast	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

[6] In fact one needs to use martingales to get things to work out correctly.

Assuming the pattern of 00 followed by 11 repeats indefinitely, the long run frequency of 0 is 0.5 as anticipated by the forecast. In the odd periods, the long run frequency of 0 is also, as anticipated, 0.5. In the even periods, it is also 0.5. However, if the probability of 0 in every period were, in fact, 0.5 then we would expect that the frequency of 0, after 00 was observed, to be 0.5. In the data, this frequency is zero. Analogously, the frequency of 0 in period $4n+1$, n a natural number, should be 0.5 when it is one.

Thus, dividing the sequence into just two subsequences is not enough. How many subsequences would suffice? To answer this question, we formalize the notion of dividing up the entire sequence of observations into subsequences.[7]

Imagine a rule that, at the end of each period, decides whether or not to mark the period (as a function of the past). The marked periods define a subsequence on which the forecasts (made in those periods) could be compared to the outcomes realized next period. One rule might be to mark every even numbered period. The forecasts made in the even periods will be compared with the outcomes realized next period. Another rule would be to mark the period if the current outcome is 0. The forecasts made in the periods that 0 occurred will be compared with the outcomes realized next period.

A rule that decides which periods to mark (as a function of past and current outcomes) is called an *outcome-based checking rule*. Formally, an outcome-based checking rule is a function from finite sequences of outcomes to $\{0,1\}$. We say that the rule is active when it assumes the value 1 for that period. The marked periods are those in which the rule is active. An outcome-based checking rule could be active when the last three observation were 010, when the period is a prime-number, etc.

Outcome-based checking rules mark a period based on past and current outcomes only. However, if forecasts change then we may want a checking rule that marks a period as a function of the forecasts as well. Fix an outcome-based checking rule and an interval D of possible forecasts. An associated forecast-based checking rule will mark those periods marked by the outcome-based checking rule *and* when the forecast lies in D . That is, a forecast-based checking rule is active when the outcome-based checking rule is active *and* the forecast is within some interval (these intervals form a partition of $[0,1]$). For example, consider the outcome-based checking rule that is active in the even periods. Consider the partition [0, 0.5) and [0.5, 1]. A forecast-based checking rule (associated with this outcome-based checking rule) is active in the even periods when the forecast for 1 is less than 0.5. Another forecast-based checking rule is active in the even periods when the forecasts for 0 is greater than 0.5. For each forecast-based checking rule, there is an associated subsequence of active periods. The forecasts will be compared to the data separately in each of these subsequences.

Given a collection $\mathcal{C}$ of outcome-based checking rules and a partition of $[0,1]$, we say that a sequence of forecasts is *calibrated* with respect to the observed data if the average forecasts match empirical frequencies, in the subsequence specified by the forecast-based checking rule associated with the outcome-based checking rules in $\mathcal{C}$. Informally, a sequence of forecasts is *calibrated* if, in the subsequences specified by $\mathcal{C}$, the frequency of 0 is p in the sub-subquences in which the forecast is p.

[7]This is an idea due to Dawid [6]. See also [18].

The examples above show that forecasts matching empirical frequencies for *finitely* many checking rules may fail to capture relatively simple patterns. However, consider a *countable* collection of outcome-based checking rules that include all functions (mapping finite sequences of outcomes to $\{0,1\}$) implementable by a recursive algorithmic. Consider a countable partition of $[0,1]$. This collection of forecast-based checking rules is also countable. Notice that the countable collection of checking rules we focus on are all rules that can be implemented by a Turing machine. If the forecasts match the empirical frequencies for all these forecast-based checking rules then no comparison between frequencies and the forecasts, that is implementable by a Turing Machine, would reject the hypothesis that the forecasts are correct.

The main result of Sandroni, Smorodinsky and Vohra [32] shows that, given any countable collection of outcome-based checking rules and countable partition of the entire interval, there is a forecasting scheme that generates sequences of calibrated forecasts on *every* possible infinite string of data. So, if a forecaster uses this forecasting scheme then after some point in the future, when he looks backwards, he will always see that the time average of the forecasts are close to the empirical frequencies. In this sense, he will not see a contradiction between the forecasts and the data.[8]

A forecast that would be calibrated with respect to all checking rules (not just countably many) would satisfy the stronger property of merging (see [18]). The distinction between the countable and the uncountable case highlights the weakness of the calibration criterion. Calibration is a guarantee that at some distant point in the future, looking *back*, the forecast will be consistent with past outcomes. Merging is a guarantee that at some distant point in the future, looking *forward*, the forecast will be consistent with future outcomes.

4 Testing

Are there tests, other than calibration, that can distinguish between a forecaster who knows the underlying distribution of the process being forecast from one who 'games' the test? Rather than run through a long collection of criteria, we follow Sandroni [32] and focus on properties that such tests should have.

Formally, a test takes as input a forecasting algorithm, a sequence of outcomes and after some period accepts the forecast (PASS) or rejects it (FAIL). Two properties that such a test should possess appear compelling. First, the test should declare PASS/FAIL after a finite number of periods. This seems unavoidable for a practical test. Second, suppose the forecast is indeed correct i.e., accurately gives the probability of a state being realized in each round. Then, the test should declare PASS with high probability. Call this second condition "passing the truth." In other words, the probability of a type I error should be small.[9]

Call a test that satisfies these two conditions a *good* test. A test based on calibration is an example of a good test. A forecaster with no knowledge of the underlying distribution that can

[8] Lehrer [20] establishes this for the special case of outcome based checking rules only. See also Vovk and Shafer [36].

[9] A type I error occurs when the true hypothesis is rejected. Acceptance of a false hypothesis is a type II error. Not knowing which is which is a type III error.

pass a good test with high probability on all sequences is said to *ignorantly* pass the test. Implicit in the notions defined is that the forecaster knows the test.

To define these notions precisely we require some notation. A sequence $s \in S^n$ and a forecasting algorithm F determine a probability measure $\bar{F}^s$ on $[0,1]^n$, where conditional on (s^r, f^{r-1}), the probabilities of forecasts next period are given by $F(s^r, f^{r-1})$. The vector of realized forecasts associated with F on a sequence s will be denoted $f(s)$.

Denote the unknown data generating process by P. Given P and $s^r \in [0,1]^r$ let $P_{s^r} \in [0,1]$ be the probability that $s_{r+1} = 1$ conditional on s^r. Given P let $F^P(s) \in [0,1]^n$ be the forecast sequence such that $f_r^P(s) = P_{s^r}$.

A *finite test* is a function $T : S^n \times [0,1]^n \to \{0,1\}$. After a history of n forecasts and outcomes are observed, a test must either accept (PASS) or reject (FAIL) the forecast. When the test returns a 0 the test is said to fail the forecast based on the outcome sequence. When the test returns a 1 the test is said to PASS the forecast based on the outcome sequence.

One can think of a finite test as a subset of $S^n \times [0,1]^n$. If the history of n forecasts and outcomes lies in this subset, the forecast is failed, otherwise it is passed. A non-finite test would be a subset of $S^*t \times [0,1]^*$, where $[0,1]^*$ is the set of infinite sequences of numbers in $[0,1]$. Call the set of outcomes on which a forecast would be rejected by a test T, T's *rejection set*.

A test is said to pass the truth with probability $1-\epsilon$ if

$$\Pr_P(\{s : T(s, F^P(s)) = 1\}) \geq 1 - \epsilon$$

for all P.

A test T can be *ignorantly* passed by a forecasting algorithm F with probability $1-\epsilon$ if for every $s \in S^n$,

$$\Pr_{\bar{F}^s}(\{f : T(s, f(s)) = 1\}) \geq 1 - \epsilon.$$

Hence, F can ignorantly pass T if on any sequence of outcomes, the realized forecast sequence will be passed with probability at least $1-\epsilon$ (under the distribution induced by the forecasting scheme). A test T is said to fail the forecasting algorithm F on the distribution Q with probability $1-\epsilon$ if

$$\Pr_Q(\{s : \Pr_{\bar{F}^s}(\{T(s, f(s)) = 1\}) \geq 1 - \epsilon\}) \leq \epsilon.$$

For every good test, Sandroni [32] shows there exists a randomized forecasting algorithm that will ignorantly pass the test.

Theorem 1 *Suppose a finite test T passes the truth with probability $1-\epsilon$. Then, there is a forecasting algorithm F that can ignorantly pass T with probability $1-\epsilon$.*

Therefore, no good test can distinguish between a forecaster who knows the underlying distribution of the process being forecast from one who 'games' the test. In this sense, Sandroni's Theorem is an impossibility result. The argument is similar to the one that established the existence of a calibrated forecast except it requires a stronger minimax theorem. One could imagine 'secret' tests

that are not revealed to the forecaster. A forecaster faced with a secret test can always treat this as going up against a mixture of known tests. A mixture over good tests is also a good test.[10]

One way to see why Sandroni's impossibility result holds is to consider a good test that must decide after one period. Given it must pass the truth, it has little choice but to pass all forecasters. Now, what about n periods? The decision to pass or fail the forecast is based on some n-period sequence, s^n. Unfortunately, there are uncountably many infinite sequences, s^*, which contain s^n as a prefix. Thus, two forecasts that agree on the first n observations need not agree subsequently. This makes it difficult for a good test to distinguish between a correct forecast and one that 'games' the test.[11]

The deeper reason lies in the logic of the minimax theorem. Suppose, nature had a mixed strategy that the forecaster could not 'beat'. Then, nature could just announce this strategy, i.e., announce the distribution itself. But, the forecaster could just use this distribution to forecast!

It is natural to ask if a test, using a proper scoring rule like *log-loss*, can circumvent the difficulties identified by Sandroni's result. Here one penalizes the forecaster $\log p$ if the forecaster predicts a probability p of rain and it rains and a penalty of $\log(1-p)$ if it doesn't rain. The lowest possible score that can be obtained is the long-run average entropy of the actual distribution governing the frequency of rain. One could imagine a test passing the forecaster if its log loss matches the entropy. However, such a test would need to know the entropy of the distribution. As noted in the introduction, we are concerned with tests which operate without any prior knowledge of the distribution. Proper scoring rules are good methods to compare two forecasters but are not useful for testing the validity of a forecaster against an unknown distribution of nature.

If one replaces the proletarian term forecast by the more aristocratic, theory, Sandroni's impossibility result is a strike against the idea that a theory can be verified on purely empirical grounds. More generally it is a criticism of the classical notion of induction: the ability to reason about the future from the past.[12]

4.1 Surmounting The Impossibility

Any impossibility theorem can be breached by relaxing at least one of its assumptions. Such is the case here. Technically it amounts to identifying conditions under which the minimax theorem fails.

Dekel and Feinberg [7] surmount the impossibility by dropping the requirement that the test be finite. The test in [7] takes as input the forecasting algorithm itself rather than just the realizations. Because a forecasting algorithm specifies a conditional probability given any history, it essentially specifies a distribution, μ over S^*. Interpret S^* to be the set of binary expansions of numbers in

[10] Mixtures over tests that are not good can also be accommodated as long as the probability assigned to 'not good' tests is not to large.

[11] This echoes Sextus Empiricus' objection to generalizing from a finite collection of the particulars because *all* of the particulars are infinite in number.

[12] There is also a connection to the modern formulation of the problem of induction due to Goodman [14] which we recount. First, call a thing *grue* if and only if it has been observed to be green before a finite time t or blue after that time. Recall now that all emeralds ever seen are both green and grue. Why is it, Goodman asks, that we believe that after after time t we will find green but not grue emeralds? Goodman argues that an appeal to Occam's razor does not apply here.

$[0,1]$. Thus, the forecaster specifies a measure μ over $[0,1]$ and nature picks an element in $[0,1]$. Using only μ and the lone element in $[0,1]$ the test must decide whether to pass or fail the forecast. Framed this way, the testing question appears unnatural since one must decide based on a 'single draw' whether to pass or fail the forecaster. It seems very likely that any test that passes the truth with high probability can be ignorantly passed. [13]

Dekel and Feinberg [7] show that associated with every distribution μ over S^* is a 'small' set $K_\mu \subset S^*$ with two properties.

1. $\mu(K_\mu) = 1$.

2. The set of distributions that assign K_μ positive probability is also 'small'.

The intuitive idea is that to every distribution one can assign an essentially unique signature that is hard to duplicate by another distribution. The notion of 'small' used here is that of category. The set K_μ is category 1 (countable union of nowhere dense sets) and the set of distributions that assign positive probability to K_μ is also category 1. The test fails the forecast if the outcome falls outside of K_μ and passes the forecast otherwise.

The topological notion of 'small' differs from the measure theoretic notion of small.[14] That difference is exploited in Olszewski and Sandroni [26] to show that the test in [7] can be ignorantly passed by a suitable randomized forecasting scheme. However, as shown in [7], assuming the continuum hypothesis, the test in [7] can be modified so that it cannot be ignorantly passed. Furthermore, the set of outcomes on which an ignorant forecaster would fail is uncountable.

Reliance on the continuum hypothesis is problematic. Olszewski and Sandroni [25] describe a test that cannot be ignorantly passed that does not rely on the continuum hypothesis. Instead, it invokes the axiom of choice (AC).

Olszewski and Sandroni [26] bypass the continuum hypothesis by requiring the test to declare FAIL in a finite number of periods but PASS 'at infinity'. This is consistent with Popper's notion of falsifiability. They show the existence of a test that passes the truth with high probability that cannot be ignorantly passed. However, the number of periods before an ignorant forecaster is failed can be extremely large and depends on the forecaster.

Olszewski and Sandroni [27] observe that the tests considered in [7] and [25] rely on the forecasting algorithm itself. Specifically, the test can use the predictions the forecast would have made along sequences that did not materialize. As noted in [27] this is not the case for many natural tests. For this reason they restrict attention to tests that are not permitted to make use of counterfactual predictions. Essentially, two different forecasting algorithms that produce the same forecast on a realization must be treated in the same way. The test must declare FAIL in a finite number of periods but can PASS 'at infinity'. Under these conditions they recover the impossibility result. Specifically, if such tests pass the truth with high probability they show that for each such test, there is a forecasting algorithm that can ignorantly pass the test. Shmaya (2008) [35] shows that

[13] In fact, even if we allow for $k > 1$, independent draws from $[0,1]$, the challenge is the same. One can interleave the k independent binary expansions into a single sequence.

[14] See problem 55(d) in [31].

one can relax the condition that the test must declare FAIL in finite time and recovers the same impossibility result. Crucial to his proof is that Blackwell games are determined (see [22]).

The result in [35] suggests attention be directed to the axiom of determinacy (see Mycielski and Steinhaus [24]). The axiom presumes a certain class of extensive form two person games of perfect information with a countable number of moves on each path is determined. The game has players who take turns choosing the next term in the decimal expansion of a number in $[0, 1]$. If the number thus chosen lies in some given $A \subseteq [0, 1]$, player 1 wins, otherwise player 2 wins.

The axiom of determinacy (AD) is incompatible with the axiom of choice (AC). However, (AD) like (AC), is consistent with Zermelo-Fraenkel set theory.[15] We **conjecture** that the existence (or not) of tests that cannot be ignorantly passed hinges on which of (AC) or (AD) one accepts.

This results of [27] and [35] appear to dash any hopes of a purely empirical approach to validating a forecast. The papers, summarized next, suggest that it is too early to muffle the drums and call out the mourners. These papers take one of two approaches. The first impose complexity constraints on the test as well as the forecaster.

Most practical tests, for example, have a complexity that is polynomial in the length of the history, so it seems reasonable to restrict attention to good tests that have a complexity that is polynomial in the length of the history. Restricting the test in this way should, make it 'easier' to be ignorantly passed. It seems natural to conjecture that for every polynomial time test that passes the truth with high probability, there exists a polynomial time randomized forecasting algorithm that will ignorantly pass the test. This is not the case. Fortnow and Vohra (2009) [9] describe a linear time test that can be ignorantly passed only if the the forecaster were able factor numbers under a specific distribution. The existence of an efficient (i.e. probabilistic polynomial time) algorithm for factoring composite numbers is considered unlikely. Indeed, many commercial available cryptographic schemes are based on just this premise. This result suggests that the 'ignorant' forecaster of Sandroni [32] must have a complexity at least exponential in n. Hence, the 'ignorant' forecaster must be significantly more complex than the test. In particular its complexity may depend on the complexity of nature's distribution.

The idea behind this result is to interpret the observed sequence of 0-1's as encoding a number followed by a list of its possible factors. Call a sequence correct, if its suffix is a correct factorization of its prefix. The test fails any forecaster that does not assign high probability to these correct sequences when they are realized. Consider now the distribution that puts most of its weight on a correct sequence. If the forecaster can ignorantly pass the test, it must be able to identify correct sequences.[16]

The second approach (related to the above) relaxes the condition that the Type I error must be small. One way to do this is to restrict 'nature' to picking it's distribution from a restricted set known to the test. In doing so we step away from a test that is pure empiricism, since the test incorporates prior knowledge. This prior knowledge amounts to a restriction on the class of forecasts considered. The test can simply fail any forecast that not in this class. Observe that doing so raises the probability of a type I error.

[15] Many games can be shown to be determined without an appeal to (AD). Blackwell games are an example[22].

[16] Huh and Shmaya [16], in the same vein, suppose that forecaster and test must be Turing computable.

Clearly, how one restricts the forecaster (or nature) matters. [26] show that when nature is restricted to picking distributions from a certain non-convex set, there exists a test that cannot be ignorantly passed. The restriction is no more than a counter-example to a possible generalization of their main result. Al-Najjar, Sandroni, Smorodinsky and Weinstein [2], propose two criteria for identifying a 'natural' restriction.

1. Learnable: Nature's distributions should permit the forecaster to learn from data.
2. Predictive: The forecaster should not need to keep learning forever; eventually, she will have learned enough so that new evidence will have a small effect on predictions about the distant future.

The notions are formalized in Jackson, Kalai, and Smorodinsky [17]. Restricting nature to distributions that are learnable and predictive, [2] design a test in which the forecaster is required to submit a date, d, by which she will have learned enough to deliver sharp predictions about future frequencies. They show this test passes a forecaster who knows the data-generating process and cannot be passed by an uninformed forecaster (restricted to forecasts in this class).

The difficulty that [2] overcome with this restriction is that a distribution can be represented as a convex combination of 'component' distributions in many ways. As an example, nature first draws a number, θ, uniformly at random from $[0, 1]$. Next, nature generates a 0-1 sequence by flipping a coin that will come up heads ($= 1$) with probability θ. Here is a second representation. Imagine two coins, called *high* and *low*. The *high* coin comes heads with probability θ^h, where θ^h is a draw from the uniform distribution over $[0.5, 1]$. The *low* coin comes heads with probability θ^l, where θ^l is a draw from the uniform distribution over $[0, 0.5]$. This observation suggests an alternative restriction on Nature's distribution: that the set of distributions Nature may employ is suitably non-convex. Lambert [19] takes such an approach as well as providing examples of natural instances that satisfy his notion of non-convexity.[17]

5 Multiple Forecasters

Rarely is it the case that a single theory or forecast is subject to an up or down decision. Rather, theories and forecasts are compared and the best of the lot is picked. At first blush, this makes the work just summarized irrelevant. Not so. Imagine one is being compared against another forecaster, call them C. Now suppose, your forecasts will be compared with C's forecast in some way and, eventually, one of you will be selected. Suppose also, you know both C's forecasting algorithm as well as the metric by which you will be compared with C. Then, C's forecasts and the metric constitute a test and the previous results apply. They apply because you knew both the metric and C's forecasts and therefore knew the test. In some contexts, it is unreasonable to expect that you would know the forecasting algorithm of the competing forecaster. In this case, one is faced with a 'secret' test in the sense that you cannot tell ahead of time what you will be 'tested' on. As there is a possibility a 'secret' test may fail the truth, one may wonder if amongst the alternative forecasts being evaluated, if there is one that is 'correct', could a 'secret' test determine it? Yes.

[17] The main result is a good test that can be passed if and only if the forecasts merge with the true distribution.

Feinberg and Stewart [8], for example, propose a cross-calibration test of predictions by multiple potential forecasters. The test checks whether each forecaster is calibrated conditional on the predictions made by other forecasters. They show this test is a good test that cannot be ignorantly passed.

Al-Najjar and Weinstein [1] show that a simple 'reputation-style' test can distinguish between two experts one of whom is informed about the true distribution.[18] The test presumes no prior knowledge of the true distribution, achieves any desired degree of precision in some fixed finite time, and does not use 'counterfactual' predictions. It exploits a rate of convergence of supermartingales result.

Olszewski and Sandroni [28] also consider the case of multiple forecasters but do not assume that amongst them is one that 'knows' the truth. Assume a test that will compare the forecasters and select one if it knows the truth. Suppose none of the forecasters knows the truth. Then, they can still independently produce false forecasts that will pass the test, independently of how the data evolve.

6 Conclusion

Looking into the future, we see three lines of inquiry as worth pursuing. The first is to see how far the calibration criterion can be used in place of the Bayesian assumption. One example of just such a substitution is in the connection between calibration and correlated equilibrium (see [12]). One can imagine others, for example, no trade theorems. The second, is how to choose amongst different forecasters when what one cares about is not the forecast itself but its payoff implications (see [29]). The third, is understanding the relationship between the work described and the problem of generating pseudo-random sequences. Checking whether a sequence is random is in a sense 'dual' to the problem of verifying a probability forecast. Instead of being given a sequence and coming up with a distribution, we are given a distribution and must come up with a sequence that appears as if it could be generated by the given distribution.

Acknowledgements

We thank Wojciech Olszeski, Mallesh Pai, Alvaro Sandroni, Itai Sher and Lance Fortnow for useful comments.

References

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

[2] Al-Najjar, N., A. Sandroni, R. Smorodinsky and J. Weinstein (2009) "Testing Theories with Learnable and Predictive Representations," manuscript.

[18] The test can be interpreted as a likelihood ratio test.

[3] Brier, G. (1950) "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, 78, 13.

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

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

[6] Dawid, A.P. (1985) "Calibration-based Empirical Probability" (with Discussion). *Ann Statistics* **13**, 1251-1285. Reprinted in *Probability Concepts, Dialogue and Beliefs*, edited by O. F. Hamouda and J. C. R. Rowley. Edward Elgar Publishing Ltd (1997), 174-208.

[7] Dekel, E. and Y. Feinberg (2006) "Non-Bayesian testing of a Stochastic Prediction," *Review of Economic Studies*, 73, 893 - 906.

[8] Feinberg, Y. and C. Stewart (2008) "Testing Multiple Forecasters," *Econometrica* 76 (3), 561582.

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

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

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

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

[13] Fudenberg, D. and D. Levine (1999b) "An Easier Way to Calibrate," *Games and Economic Behavior* **29**, 131–137.

[14] Goodman, Nelson (1955). Fact, Fiction, and Forecast. Cambridge, MA: Harvard University Press.

[15] Hart, S. and A. Mas-Colell (2001) "A General Class of Adaptative Strategies," *Journal of Economic Theory*, **98**, 26-54.

[16] Huh, T-W. and E. Shmaya (2010). "Expressible tests need not be manipulable," manuscript.

[17] Jackson, M. O., E. Kalai, and R. Smorodinsky (1999) "Bayesian Representation of Stochastic Processes under Learning: de Finetti Revisited," *Econometrica*, 67(4), 875893.

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

[19] Lambert, N. (2010), personal communication.

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

[21] Mannor, S. and G. Stoltz (2009) "A Geometric Proof of Calibration," manuscript.

[22] Martin, D. A. (1998) "The determinacy of Blackwell games," *Journal of Symbolic Logic*, 63,15651581.

[23] Murphy, A.H. (1973) "A new vector partition of the probability score," *Journal of Applied Meteorology* 12 (4), 595600.

[24] Mycielski, J. and Steinhaus, H. (1962) "A mathematical axiom contradicting the axiom of choice," *Bulletin de l'Acadmie Polonaise des Sciences. Srie des Sciences Mathmatiques, Astronomiques et Physiques* , 10, 13.

[25] Olszewski, W. and A. Sandroni (2008) "A nonmanipulable test," *Annals of Statistics*, 37(2), 1013-1039.

[26] Olszewski, W. and A. Sandroni (2009) "Strategic Manipulation of Empirical Tests," *Mathematics of Operations Research*, 34(1), 57-70 .

[27] Olszewski, W. and A. Sandroni (2006) "Counter Factual Predictions," manuscript, Northwestern University.

[28] Olszewski, W. and A. Sandroni (2009) "Manipulability of Comparative Tests," *Proceedings of the National Academy of Sciences, USA*, 31, 5029-5034, 2009.

[29] Olszewski, W. and M. Peski (2009) "The principal-agent approach to testing experts," mansucript.

[30] Page, L. and R. T. Clemen (2010) "Do prediction markets produce well calibrated probability forecasts?," manuscript.

[31] Royden, H. L. (1988) Real Analysis, Prentice-Hall, Third edition.

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

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

[34] Schervish, M. (1985) "Comment on paper by Oakes," *Journal of The American Statistical Association*, **80**, 341-342.

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

[36] V. Vovk and G. Shafer. (2005) "Good randomized sequential probability forecasting is always possible," *J. Royal Statistical Society: Series B*, 67(5), 747-763.

[37] Tetlock, P. (2005) Expert Political Judgement: How Good Is It? How Can We Know?, Princeton University Press.

[38] Tetlock, P. (2006) "How Accurate Are Your Pet Pundits?," http://www.project-syndicate.org/commentary/tetlock1/English.