**Research Paper No. 1957**

# Testing Multiple Forecasters

**Yossi Feinberg**
**Colin Stewart**

**January 2007**

RESEARCH PAPER SERIES

STANFORD
GRADUATE SCHOOL OF BUSINESS

# Testing Multiple Forecasters*

Yossi Feinberg† Colin Stewart‡

Stanford University Yale University

January 2007

**Abstract**

We consider a *cross-calibration* test of predictions by multiple potential experts in a stochastic environment. This test checks whether each expert is calibrated conditional on the predictions made by other experts. We show that this test is good in the sense that a true expert—one informed of the true distribution of the process—is guaranteed to pass the test no matter what the other potential experts do, and false experts will fail the test on all but a small (category one) set of true distributions. Furthermore, even when there is no true expert present, a test similar to cross-calibration cannot be simultaneously manipulated by multiple false experts, but at the cost of failing some true experts. In contrast, tests that allow false experts to make precise predictions can be jointly manipulated.

*We wish to thank Nabil Al-Najjar, Brendan Beare, Dean Foster, Sergiu Hart, Stephen Morris, Wojciech Olszewski, Alvaro Sandroni, Jakub Steiner, and Jonathan Weinstein for helpful comments and suggestions.

†email: yossi@gsb.stanford.edu

‡email: colin.stewart@yale.edu

# 1 Introduction

Economic and other scientific models commonly include a stochastic component. A novice tester may wish to test potential experts who each claim to possess a predictive stochastic model—a theory. Assuming the tester has no prior distribution over the stochastic process at hand, the question is whether, by simply observing a sequence of probabilistic predictions by the experts and the realization of the process, the tester can distinguish true experts from charlatans.

In this paper we provide a method for reliably testing sequential predictions in the presence of multiple potential experts. Contrary to the case of testing a single expert, with two or more potential experts we can construct a test revealing their types by pitting their predictions against one another.

In many cases, theories allow only for sequential predictions in the sense that they can only assess the probabilities of future events given some information about the realization so far. For example, a weather forecaster may have a model determining the probability of rain tomorrow based on the barometric pressure in the past week, but be unable to predict the distribution of barometric pressure tomorrow. Hence he cannot use his model to predict the probability of rain two days from now until tomorrow's barometric pressure has been realized. Similarly, an economic model might connect one observable indicator—interest rates, say—to the future distribution of another indicator—say unemployment—yet be completely silent on the future distribution of interest rates. In this sequential setting, a novice tester must wait for the realization to unfold before obtaining the potential expert's prediction. He cannot ask the weather forecaster for the ex-ante distribution of rain from now to eternity.

A most intuitive sequential test asks that the expert predictions be calibrated, i.e. that the empirical distribution conditional on his prediction converge to that prediction. For example, if the expert states that the probability of an increase in unemployment is 40%, we would like to see that, on average, the unemployment rate rose 40% of the time in those periods for which this 40% prediction was made. Dawid (1982,1985) proposed

this test and showed that an expert predicting according to the true distribution of the process will be calibrated in this sense. However, Foster and Vohra (1988) demonstrated that this test can be manipulated by a false expert: there exists a mixed forecasting strategy that is calibrated with probability one on *every* realization of the process.

This negative result has been extensively generalized to many other classes of tests, see Kalai, Lehrer, and Smorodinsky (1999); Fudenberg and Levine (1999); Lehrer (2001); Sandroni, Smorodinsky, and Vohra (2003); Sandroni (2003); and Vovk and Shafer (2005). Recently, Olszewski and Sandroni (2006a) obtained the strong result that *all* sequential tests[1] of a single potential expert can be manipulated. These results stand in sharp contrast to the case of *ex ante* predictions. If the tester can ask the potential expert to predict the entire distribution of the process on day one, then there exists a good test that cannot be manipulated, as first shown by Dekel and Feinberg (2006); see also Olszewski and Sandroni (2006b) for a stronger result.

We show that, with more than one potential expert, the situation is very different: there is a good sequential test that cannot be manipulated by false experts. In fact, this test is a simple extension of the calibration test. The crux of the matter is that with more than one expert, there is more information for the tester, even if the information is provided by a false expert. The tester requires not only that predictions be verified against the realization, but also that they be verified against the predictions made by the other potential experts. If the potential experts are unable to correlate their actions, then the ability to manipulate is dramatically reduced.

We call our sequential test the *cross-calibration test*. This test compares the empirical frequencies of events conditional on the *joint* predictions made by the experts. For example, consider all of the periods where one potential expert forecasts the probability of increase in unemployment to be 40%, while another potential expert puts it at 20%. Conditional on these predictions, the empirical distribution cannot be both 40% *and* 20%. Hence, if such a disagreement in predictions occurs infinitely often, we are guaranteed that at least one of the potential experts will not be calibrated with

[1]They refer to these tests as those that do not make use of counterfactual future predictions.

respect to this test.

We show that a true expert predicting according to a model based on a distribution $P$ is guaranteed to pass the cross-calibration test with $P$-probability one. In other words, if $P$ indeed governs the process, a true expert is bound to be well cross-calibrated no matter what strategy—pure or mixed—is employed by the other potential experts. Note that the true expert need not know the entire description of the distribution $P$—he only needs to know the conditional probability in each period given the realized history so far. We also show that the classic calibration test of a single forecaster is a good test (as defined by Dekel and Feinberg (2006)) in the sense that the set of realizations on which a model following $P$ will be calibrated is a category one set—a countable union of nowhere dense sets. In particular, cross-calibration is also a good test since it has no Type I errors—a true expert is guaranteed to pass, and the set of realizations on which he passes is a subset of those on which he is calibrated when tested in isolation.

Unlike tests of a single expert, with multiple experts, the test of each one generally depends on the forecasts made by others. For a true expert, the strategies of the other potential experts are irrelevant, but for a false expert they are not. Thus false experts may attempt to tailor their strategies based on their knowledge of the strategies of others.

Before describing the results in detail, let us state more precisely what we mean by "knowing" a distribution. Consider the case of a single potential expert. We can view the test as a game in which nature chooses a realization according to a mixed strategy $P$. The expert is required to make predictions, either *ex ante* if a true expert is supposed to possess knowledge of the full distribution $P$, or sequentially if a true expert is supposed to have a model that reveals Nature's "behavioral strategy"—the conditional probability generated by $P$ given the *particular* history realized so far. In either case, a true expert is informed (at least partially) of Nature's strategy. In contrast, a false expert has no such information. In particular, Nature's strategy is not part of an equilibrium of the game, and randomization by Nature is not defined

in terms of the expert's conjecture (or uncertainty) regarding Nature's pure strategy. The property we wish to test is whether or not each potential expert knows Nature's mixed strategy (or the relevant conditionals in the sequential case).

Consider a false expert being tested alongside a true expert. If the false expert knew the true expert's strategy, then he would himself possess the same knowledge as the true expert, making him a true expert rather than a false one. Thus knowing the *strategy* of another true expert amounts to *being* a true expert oneself. As in the single expert case, a test of multiple potential experts can only distinguish their knowledge of the distribution if it cannot be manipulated by a false expert. The question is therefore whether there exists a (mixed) strategy for the false expert that is sure to pass the test regardless of the true distribution $P$, in other words, without any knowledge of the true expert's strategy. The answer is no. In fact, a false expert is guaranteed to fail the test for most distributions $P$. When a true expert is present, we show that, for every (mixed) forecasting strategy, a false expert using this strategy passes the cross-calibration test with probability zero except on a category one set of true distributions $P$. Thus a false expert can only be guaranteed to pass the test for a small set of strategies by Nature when these strategies determine the predictions made by the true expert. These results remain unchanged if more experts, true or false, are added to the mix.

In independent work, Al-Najjar and Weinstein (2006) consider a test which compares the likelihoods of predictions made by a false and a true expert. Their test is relative: the single expert who performs best is the one who passes. Al-Najjar and Weinstein elegantly show that a false expert cannot manipulate this comparative test: whichever strategy the false expert uses, there will be a true expert who will provide more accurate predictions. While their test need not pass the true expert—is not free of Type I errors[2]—it successfully identifies probabilities close to the true distribution, i.e. even if it selects the false expert the predictions that performed best are likely to

[2]For many true distributions, no matter how long the test is, there exist strategies for the false expert that make it likely that the true expert fails.

be close to the true probabilities. Moreover, these probabilities are identified for all but a uniformly bounded finite number of periods. In contrast, cross-calibration has no Type I error, and in particular always identifies the true probabilities[3], but we do not know if a uniform bound on the number of periods can be obtained. Moreover, the false expert cannot manipulate cross-calibration and is sure to fail on *most* true distributions.

If none of the potential experts is a true expert, then it may be easier for false experts to manipulate the test since they are not tested against the true conditional probabilities. The question is to what extent the false experts can manipulate the test, individually or jointly. The answer depends on what the false experts know, and requires a modification of the cross-calibration test. If one false expert knows the strategies of all other false experts, then he is facing a randomized calibration-type test. As shown by Lehrer (2001), this false expert has a strategy allowing him to manipulate the test. This result holds more generally for *any* test of multiple experts that relies only on sequential predictions. If there existed strategies for the other forecasters that prevented manipulation in this way, then we could use these strategies to design a non-manipulable sequential test of a single forecaster, which Olszewski and Sandroni (2006a) proved to be impossible.

While the manipulation by a single false expert cannot be prevented if he knows the others' strategies, in the other extreme, we know that an expert cannot manipulate the test no matter what others are doing (this follows from the case where there is a true expert among those tested). There remains the question of whether it is possible for multiple false experts to pass the test simultaneously. We answer this question for a modification of the cross-calibration test. We consider a *strict* cross-calibration test which requires the empirical frequencies to lie within the predicted intervals, not only on their boundaries. This test is no longer free of Type I errors—a true expert may fail the test—but only for knife-edge cases. For the strict test, although every expert has

[3]In our test, it is possible for a false expert to pass alongside a true expert. In order for this to occur, the predictions of the false expert must be close to the true probabilities in all but finitely many periods; the test therefore identifies probabilities close to the true ones.

a best response to the others' strategies, there is no Nash equilibrium as long as each expert always strictly prefers to pass the test. If they all know each others' strategies, they cannot all be playing best responses. Except on a small set of realizations, the probability that at least two potential experts pass the test simultaneously is zero. To summarize, for the strict test with no true expert, the false experts cannot know each others' strategies and simultaneously manipulate, nor can they manipulate the test if they do not know each other's strategies. Furthermore, even if they try to coordinate their strategies, they can only jointly pass the cross-calibration test with positive probability on a category one set of realizations. Hence, the strict cross-calibration test cannot be manipulated (in a strong sense) even when there is no true expert present.

We conclude by showing that it is necessary to place some restriction on the richness of predictions announced in the test in order to prevent joint manipulation by multiple false experts. If the experts are allowed to report exact probabilities, rather than an interval of probabilities as in the cross-calibration test, then they are able to correlate their predictions and manipulate the test simultaneously.

Our results reinforce the role of the topological notion of category one as a measure of *smallness* in a forecasting environment. We show that the standard calibration test is a good test—predictions based on a distribution $P$ pass on a category one set of realizations that has $P$-probability one. Hence, by Foster and Vohra (1998), calibration is a simple manipulable good test.[4] Moreover, the current examples of single expert tests that cannot be manipulated are good tests (see Dekel and Feinberg (2006) and Olszewski and Sandroni (2006b)). For multiple experts, cross-calibration also belongs to the class of good tests. It is also the case that false experts can jointly pass the strict cross-calibration test with positive probability on at most a category one set of realizations. What is perhaps most interesting is that when a false expert faces a true expert in the (good) cross-calibration test, he can manipulate against at most a category one set of possible *distributions*, the same notion characterizing the extent of

[4]The first example of a manipulable good test was given by Olszewski and Sandroni (2006b).

Type II errors for good tests.

## 2 The Cross-Calibration Test

The environment we consider extends the classic calibration framework to allow for multiple forecasters. Let $\Omega = \left\{(\omega_t)_{t=0,1,...} \mid \omega_t \in \{0,1\}\right\}$ denote the space of possible realizations. Fix a positive integer $n > 4$, and divide the interval $[0,1]$ into $n$ equal closed subintervals $I_1, \ldots, I_n$, so that $I_l = \left[\frac{l-1}{n}, \frac{l}{n}\right]$. All results in this paper hold when $\{0,1\}$ is replaced by any finite set $S$, and the partition of $[0,1]$ is replaced by a fine enough finite partition of the distributions over $S$ into convex subsets.

At the beginning of each period $t = 0, 1, ...$, all forecasters $j \in \{1, \ldots, M\}$ simultaneously announce predictions $I_t^j \in \{I_1, \ldots, I_n\}$, which are interpreted as probabilities with which the realization 1 will occur in that period. We assume that forecasters observe both the realized outcome and the predictions of the other players at the end of each period. A (mixed) strategy for a forecaster $i$ is therefore a collection $\mu^i = \left\{\mu_t^i\right\}_{t=0}^{\infty}$ of functions

$$\mu_t^i : \{0,1\}^t \times_{j=1}^{M} \{I_1, \ldots, I_n\}^t \longrightarrow \Delta(\{I_1, ..., I_n\}),$$

where $\Delta(X)$ denotes the space of distributions over a set $X$. A strategy profile is denoted by $\mu = (\mu^1, ..., \mu^M)$.

The realization in $\Omega$ may be determined by a stochastic process. By Kolmogorov's extension theorem, a distribution $P$ in $\Delta(\{0,1\}^\infty)$ corresponds to a collection of functions

$$p_t : \{0,1\}^t \longrightarrow \Delta(\{0,1\})$$

which we also denote by $P = \{p_t\}_{t=0}^{\infty}$. Hence $P$ corresponds to a *pure* strategy that is independent of the previous predictions made by the potential experts.

The cross-calibration test is defined over outcomes $\left(\omega_t, I_t^1, \ldots, I_t^M\right)_{t=0}^{\infty}$, which specify, for each period $t$, the realization in $\{0,1\}$, together with the prediction intervals announced by each of the $M$ forecasters. Given any such outcome and any $M$-tuple

$l = \left(l^1, \ldots, l^M\right) \in \{1, \ldots, n\}^M$, define

$$\zeta_t^l = \mathbb{1}_{I_t^j = I_{l^j} \forall j = 1, \ldots, M}$$

and

$$\nu_T^l = \sum_{t=0}^{T} \zeta_t^l, \tag{1}$$

which represents the number of times that the forecast profile $l$ is chosen up to time $T$. For $\nu_T^l > 0$, the frequency $f_T^l$ of realizations conditional on this forecast profile is given by

$$f_T^l = \frac{1}{\nu_T^l} \sum_{t=0}^{T} \zeta_t^l \omega_t. \tag{2}$$

Forecaster $j$ *passes the cross-calibration test* at the outcome $\left(\omega_t, I_t^1, \ldots, I_t^M\right)_{t=0}^{\infty}$ if

$$\limsup_{T \to \infty} \left| f_T^l - \frac{2l^j - 1}{2n} \right| \leq \frac{1}{2n} \tag{3}$$

for every $l$ satisfying $\lim_{T \to \infty} \nu_T^l = \infty$.

In the case of a single forecaster, the cross-calibration test reduces to the classic calibration test, which checks the frequency of realizations conditional on each forecast that is made infinitely often. With multiple forecasters, the cross-calibration test checks the empirical frequencies of the realization conditional on each *profile* of forecasts that occurs infinitely often. Note that if an expert is cross-calibrated, he will also be calibrated.

We say that predictions are *close* to one another if the predicted intervals intersect, i.e. the intervals are either identical or have a common boundary. If the conditional empirical distribution lies exactly on the boundary between two intervals, then neither of these intervals can be ruled out in the cross-calibration test. We define the *strict cross-calibration test* to be the same as the cross-calibration test, except with disjoint intervals, for example of the form $\{[0, 1/n), [1/n, 2/n), ..., [(n-1)/n, 1]\}$. We now modify the inequality in (3) to a strict inequality (on one side of the interval) when needed, to reflect that the empirical distribution must converge to the appropriate interval.

A true expert may fail the strict cross-calibration test if, for example, the true distribution gives rise to independent probabilities $1/n - 1/e^t$ in each period $t$, causing the empirical distribution to converge to $1/n$ from below with probability one.

We consider two types of potential experts: true experts and false experts. A *true expert* knows the conditional probabilities, given the realization so far, of the distribution $P$ governing the stochastic process, while a *false expert* does not. Hence the true expert possesses the correct model. Formally, for each history $h_t = \left(\omega_s, I_s^1, \ldots, I_s^M\right)_{s=0}^{t-1}$, true experts follow the strategy defined by

$$\mu_t\left(h_t\right) \equiv I\left(p_t\left(\left(\omega_s\right)_{s=0}^{t-1}\right)\right),$$

where $I(p)$ denotes the interval containing $p$. If $p$ lies on the boundary between two intervals, we may assume that the lower interval is chosen. False experts have no knowledge of Nature's strategy, and no knowledge of any true expert's strategy if such an expert is present. They observe only the realization and past predictions of other experts, and are free to choose any strategy randomizing their prediction in each period. We assume that all experts know which, if any, of the other experts are true ones; however, relaxing this assumption has no impact on our results. Note that while we say that a true expert *follows* the distribution $P$, his strategy requires only that he know the conditional probabilities for the unraveling realization at hand, not the full distribution $P$.

Throughout the paper, we provide the proofs for two potential experts, with or without one being a true expert. The proofs are essentially the same for all other combinations of a finite number (greater than one) of potential experts. We frequently denote by Pr, instead of $\Pr_{\mu,P}$, the probability of events with respect to $\mu$ and $P$.

# 3 A True Expert

We begin by exploring the outcome of the cross-calibration test when some of the potential experts are indeed true experts. We show that no matter what others do, a true expert following the sequential predictions of a distribution $P$ of the stochastic process is guaranteed to pass the cross-calibration test with $P$-probability one.

**Proposition 1** *For every distribution $P$ governing the stochastic process, if a potential expert predicts according to a model that follows $P$, he is guaranteed to pass the cross-calibration test with probability 1 no matter what strategies the other potential experts use. That is, for any strategy profile $\mu = (P, \mu^2, ..., \mu^M)$, the set*

$$\left\{\omega \mid \forall l, \Pr_\mu\left(\limsup_{T\to\infty}\left|f_T^{l^1,...,l^M} - \frac{2l^1-1}{2n}\right| \leq \frac{1}{2n} \text{ or } \lim_{T\to\infty}\nu_T^l < \infty \middle| \omega\right) = 1\right\} \tag{4}$$

*has $P$-probability one.*

Recall that although the true expert is assumed to have a strategy $P$, he is only providing the conditional probabilities for the realized history. Hence, he need not know $P$ *ex ante*, but has the knowledge of the conditional probabilities according to $P$ once a history is realized.

**Proof.** The proof closely follows that of Dawid (1982).

Let $\mu = (P, \mu^2, ..., \mu^M)$, and fix $l$. Let $p_t = P(\omega_t|\omega_1, ..., \omega_{t-1})$.

Since $l$ is fixed, we omit it from the notation, for example in $\mu_t$ and $\zeta_t$ below. Letting $\beta_t = \frac{1}{\nu_t}$ if $\nu_t > 0$ and 0 otherwise, define

$$X_t = \beta_t\zeta_t(\omega_t - p_t).$$

Let $\mathcal{B}_t$, $\mathcal{D}_t$ and $\mathcal{F}_t$ denote, respectively, the Borel fields corresponding to the realizations $\{0,1\}^t$, the predictions $\{I_1, ..., I_n\}^t$ based on $P$, and the realizations $(\{I_1, ..., I_n\}^t)^{M-1}$ of the other forecasters' (possibly) randomized predictions $\mu_1^{-1}, ..., \mu_k^{-1}$. Since $\omega_t$ is conditionally independent of the realization of $\mu_t^j$ for $j \neq 1$ when conditioned on $\mathcal{B}_{t-1} \times \mathcal{D}_{t-1}$, and since the conditional expectation of $\omega_t$ equals $p_t$, we have $E(X_t \mid \mathcal{B}_{t-1} \times \mathcal{D}_{t-1}) = 0$.

The sequence of partial sums $S_T := \sum_{t=1}^{T} X_t$ therefore forms a martingale with respect to $(\mathcal{B}_T \times \mathcal{D}_T)$.

We have

$$E(X_t^2) = E\left((\beta_t\zeta_t)^2 \mathrm{var}\left(\omega_t \mid \mathcal{B}_{t-1} \times \mathcal{D}_{t-1}\right)\right) \tag{5}$$

$$\leq \frac{1}{4} E\left((\beta_t\zeta_t)^2\right), \tag{6}$$

and hence

$$E(S_T^2) \leq \frac{1}{4} E\left(\sum_{t=1}^{T} (\beta_t\zeta_t)^2\right) \tag{7}$$

since, for any $t' < t$,

$$E\left(X_{t'}X_t \mid \mathcal{B}_{t-1} \times \mathcal{D}_{t-1}\right) = X_{t'} E\left(X_t \mid \mathcal{B}_{t-1} \times \mathcal{D}_{t-1}\right) = 0.$$

Since $\beta_t\zeta_t = \frac{\zeta_t}{\nu_t}$ or $\beta_t\zeta_t = 0$, inequality (7) implies that $E(S_T^2) < \frac{\pi^2}{24}$ for all $T$. The martingale convergence theorem therefore guarantees the convergence of the sequence $(S_T)$ with $P$-probability one, and by Kronecker's lemma,

$$\beta_T \sum_{t=1}^{T} \zeta_t p_t - f_T \to 0 \quad \text{as} \quad T \to \infty$$

as long as $\nu_T \to \infty$. Since $p_t \in I_j = \left[\frac{2j-2}{2n}, \frac{2j}{2n}\right]$, we have

$$\left|\beta_T \sum_{t=1}^{T} \zeta_t p_t - \frac{2j-1}{2n}\right| \leq \frac{1}{2n}$$

for all $T$. The claim now follows by the triangle inequality and enumerating over $l$. ■

Note that if there are two or more true experts, they will make identical predictions and jointly pass the cross-calibration test.

# 4 False Experts: The Non-Manipulability of Cross-Calibration

When a false expert faces a multiple experts test, we must distinguish two cases. In the first case, one of the others being tested is indeed a true expert, who predicts according to a model that follows the true distribution of the process. The other case occurs when all of those being tested are false experts. What matters for manipulability, and significantly so, is whether the false expert knows what *strategies* the others are using. If he does, manipulation is possible, but if not, it is virtually impossible. When there is a true expert, knowing this true expert's strategy amounts to having the correct model, or in other words, being a true expert oneself. Hence a false expert, by definition, cannot know the strategy a true expert uses. This suffices to prevent manipulability in the presence of a true expert.

When there are only false experts, we might want to allow them to know each others' strategies in order to see if collusive manipulation is possible. It turns out that in the strict cross-calibration test, it is not. While a false expert who knows every other expert's strategy can find a strategy that will manipulate the test *given* the strategies of others, there is no equilibrium for these manipulative strategies. The others could benefit from deviating if they wish to pass the test. More precisely, there is no strategy profile where more than one potential expert can manipulate the strict cross-calibration test with positive probability on more than a small—category one—set of realizations.

## 4.1 Facing True Experts

Consider a false expert facing a true expert in a cross-calibration test. The true expert predicts according to the true distribution $P$. The failure probability is determined according to $P$ and the strategies employed by other potential experts. We would like to see the false expert fail with probability one according to the true distribution $P$, except perhaps for a small set of true distributions. The notion of smallness we employ

is that of a category one set of distributions, as suggested by Dekel and Feinberg (2006). We show that for every strategy (pure or mixed) of the false expert, for all but a category one set of distributions $P$, when the true expert follows $P$, the false expert will fail the cross-calibration test with probability one, no matter what strategies the other potential experts employ.

**Proposition 2** *In the presence of a true expert, for every strategy $\mu$ of a false expert, the set of distributions $P$ under which the false expert will pass the cross-calibration test with positive $(\mu, P)$-probability is a category one set of distributions in $\Delta(\Omega)$.*

**Proof.** We first prove the following lemma.

**Lemma 1** *If a forecasting strategy $\mu$ is cross-calibrated with respect to a true distribution $P$ with $(\mu, P)$-positive probability, then for every $\eta \in \left(0, \frac{1}{2}\right)$ there exists a finite history $h_T^\eta$ that occurs with positive probability such that*

$$\Pr(\mu \textit{ is close to } P \textit{ in every period following } h_T^\eta \mid h_T^\eta) \geq 1 - \eta. \tag{8}$$

**Proof of Lemma 1.** Recall that two predictions are close if they are identical or adjacent intervals. Assume by way of contradiction that no such history exists. In particular, (8) does not hold for the empty history. We can therefore find a finite time $t_0$ such that

$$\Pr\left(\exists s \leq t_0 \text{ such that } \mu \text{ is not close to } P \text{ at period } s\right) \geq \eta/2 \tag{9}$$

By the same argument, for every history $h_{t_0}$ that occurs with positive probability, there exists a period $t(h_{t_0}) > t_0$ such that

$$\Pr\left(\exists s \in (t_0, t(h_{t_0})] \text{ such that } \mu \text{ is not close to } P \text{ at period } s | h_{t_0}\right) \geq \eta/2.$$

Since the number of histories of length $t_0$ is finite, by choosing $t_1 = \max_{h_{t_0}} t(h_{t_0})$ we obtain

$$\Pr\left(\exists s \in (t_0, t_1] \text{ such that } \mu \text{ is not close to } P \text{ at period } s | h_{t_0}\right) \geq \eta/2$$

for every history $h_{t_0}$. Inductively, there is a finite $t_j$ such that

$$\Pr\left(\exists s \in (t_{j-1}, t_j] \text{ such that } \mu \text{ is not close to } P \text{ at period } s | h_{t_{j-1}}\right) \geq \eta/2. \tag{10}$$

Define the events

$$F_j = \left\{\exists s \in (t_{j-1}, t_j] \text{ such that } \mu \text{ is not close to } P \text{ at period } s\right\}. \tag{11}$$

We have

$$\Pr\left(\neg F_{j+k} | \neg F_j \cap \neg F_{j+1} \cap ... \cap \neg F_{j+k-1}\right) \leq 1 - \eta/2 \tag{12}$$

since (10) holds for every history $h_{t_{j-1}}$ that occurs with positive probability.

The event that the forecasts are close from some period onwards is the complement of the event that the forecasts are not close infinitely often (i.o.). We have

$$\begin{aligned}\Pr\left(\text{The experts are cross-calibrated}\right) &\leq \Pr\left(\neg \bigcap_{n=1}^{\infty} \bigcup_{j \geq n} F_j\right) \\ &= \Pr\left(\bigcup_{n=1}^{\infty} \bigcap_{j \geq n} \neg F_j\right) \leq \sum_{n=1}^{\infty} \Pr\left(\bigcap_{j \geq n} \neg F_j\right).\end{aligned} \tag{13}$$

For every $n$, we have from (12) that

$$\begin{aligned}\Pr\left(\bigcap_{j \geq n} \neg F_j\right) &= \Pr\left(\neg F_n\right) \Pr\left(\neg F_{n+1} | \neg F_n\right) \cdots \Pr\left(\neg F_{n+k} | \neg F_n \cap F_{n+1} \cap ... \cap F_{n+k-1}\right) \cdots \\ &\leq \left(1 - \frac{\eta}{2}\right)\left(1 - \frac{\eta}{2}\right) \cdots \to 0\end{aligned} \tag{14}$$

which, together with (13), implies that Pr(the complement of $F_j$ holds i.o.) = 0. Since for the forecasts to be cross-calibrated, their predicted intervals must be close from some point onwards, the probability that $\mu$ and $P$ are cross-calibrated must be 0, contradicting the assumption of the lemma, as required. ■

Fix $\eta < \frac{1}{2}$. By Lemma 1, if $P$ and $\mu$ are cross-calibrated with positive probability, then $P$ must satisfy (8) for at least one of the countable collection of finite histories. It suffices to show that the set of distributions that satisfy (8) for a given history is a category one set.

Given any finite history $h = \left(\omega_t, I_t^1, \ldots, I_t^M\right)_{t=0}^T$, let

$$\Omega(h) = \left\{\omega' = (\omega'_t)_{t=0,1,\ldots} \mid (\omega'_0, \ldots, \omega'_T) = (\omega_0, \ldots, \omega_T)\right\}$$

be the set of realizations consistent with $h$—the cylinder determined by $h$. The set $\Omega(h)$ is both open and closed—a *clopen* set. For every finite history $h$ and $\varepsilon \in (0,1)$, let $S(h,\varepsilon)$ be the set of distributions that assign probability at least $\varepsilon$ to $\Omega(h)$ *and* for which $h$ has the property of (8) (given $\eta$ and $\mu$). The set of distributions against which $\mu$ passes with positive probability is contained in the countable union

$$\bigcup_{\text{finite histories } h} \bigcup_{n=1}^{\infty} S(h, \varepsilon^n).$$

Thus it suffices to show that each $S(h,\varepsilon)$ is nowhere dense. We will show that each of these sets is closed and has empty interior.

To show that $S(h,\varepsilon)$ is closed, we want to construct for each $p \notin S(h,\varepsilon)$ an open neighborhood of $p$ that is disjoint from $S(h,\varepsilon)$. There are two cases to consider: either $p$ assigns probability less than $\varepsilon$ to $\Omega(h)$, or $h$ does not satisfy (8) (or both).

In the former case, consider the set $\{p' \mid p'(\Omega(h)) < \varepsilon\}$. This set contains $p$, and is open in the weak* topology since $\Omega(h)$ is clopen.

In the latter case, there exist some $\eta' > \eta$ and some period $\tau$ such that

$$\Pr\left(\mu \text{ is close to } p \text{ in every period from } T+1 \text{ to } \tau \text{ following } h\right) < 1 - \eta'. \qquad (15)$$

Each distribution $p'$ gives rise for each $t$ to an induced distribution $p'_t$ over finite histories of realizations $(\omega_0, \ldots, \omega_t)$. Assume first that, conditional on $h$, $p$ does not obtain conditional probabilities at the boundaries of the intervals $I_l$ between the periods $T+1$ and $\tau$. Consider the open set $U$ of distributions $p'$ for which $p'_\tau$ is at a distance less than $\delta$ from $p_\tau$ for every history. Since $p_\tau$ does not induce conditional probabilities exactly at the boundaries, we can find $\delta > 0$ sufficiently small such that $\mu$ is close to $p'$ if and only if it is close to $p$ along these finite histories. Thus (15) holds with $p'$ in place of $p$ when the probability is with respect to the original distribution $p$. Since the $p'$-probability of any event determined by time $\tau$ is within $\delta$ of its $p$-probability, (15) also holds for $p'$ with $p'$-probability when $\delta$ is sufficiently small. This guarantees that $U$ is disjoint from $S(h, \varepsilon)$. Note that $U$ is open since the probabilities that $p' \in U$ obtains are close to those that $p$ obtains on the clopen cylinders.

In the case where $p$ takes one out of a finite number of values (the boundaries of intervals $I_l$) at one out of the countable number of finite histories, we have a countable union of such probability distributions. The set of distributions taking a specific value conditional on a specific finite history $\tilde{h}$ that is obtained with positive probability is a category one set of distributions. This follows from observing that the set $S_n$ of distributions assigning at least probability $\varepsilon^n$ to $\tilde{h}$ is closed, and so is the subset of $S_n$ consisting of those distributions that assign a given probability conditional on $\tilde{h}$. Let $A$ denote the event that the history $\tilde{h} = \left(\tilde{h}_1, \ldots, \tilde{h}_t\right)$ occurs, and $B$ the event that $\omega_{t+1} = 1$. Consider the function $f_{B|A}$ that obtains the value $\alpha$ on the complement of $A$, takes the value 1 on the event $A \cap B$, and the value 0 on the event $A \cap \neg B$. When $A$ and $B$ are clopen, the set of distributions $S$ for which the integral of $f_{B|A}$ is $\alpha$ contains all distributions that assign 0 probability to $A$, and all that assign probability exactly $\alpha$ to $B$ conditional on $A$. Intersecting $S$ with the closed set $S_n$ that assigns probability at least $\varepsilon^n$ to $A$ yields a closed set $S'_n$ as required. The set $S'_n$ is nowhere dense since one can always approach any element with a sequence of distributions such that the probability of $B$ conditional on $A$ converges to, but never equals, $\alpha$. Therefore, the set of distributions with at least one conditional probability value on the boundary is

category one. Hence we can add this set to the union of the sets $S(h,\varepsilon)$ with no impact on the claim.

Finally, we must show that the interior of $S(h,\varepsilon)$ is empty. Fix $p \in S(h,\varepsilon)$. We want to construct a sequence $q^1, q^2, \ldots$ converging to $p$ such that $q^n \notin S(h,\varepsilon)$ for all $n$. As above, let $T$ be the length of the history $h$, and let $p_t$ denote the distribution over outcomes in period $t$ given the history under $p$. Define

$$q_t^n(\cdot) = \begin{cases} p_t(\cdot) & \text{if } t \leq T+n \\ 1 - \left\lfloor p_t(\cdot) + \frac{1}{2} \right\rfloor & \text{otherwise,} \end{cases}$$

where the function $\lfloor x \rfloor$ produces the largest integer not greater than $x$. Since $\eta < \frac{1}{2}$, $\mu$ cannot be close to both $p$ and $q^n$ in any period after $T+n$ with probability at least $1-\eta$. Therefore, $q^n \notin S(h,\varepsilon)$ and the proof of the proposition is complete. ■

We have shown that a false expert cannot manipulate the cross-calibration test. For any strategy he might employ, there exists a large set of distributions $P$ for each of which there exists a set of realizations having $P$-probability one on which the false expert is guaranteed to fail the test. This set of distributions contains all but a category one set in $\Delta(\Omega)$ endowed with the weak$^*$ topology.

We note that our proof implies that a false expert can be distinguished from most true experts in finite time with high probability. Consider a mixed strategy $\mu$ and a true distribution $P$ such that the probability that $\mu$ is cross-calibrated against the true expert is zero. This means that there exists some $\varepsilon > 0$ such that, for the false expert, for some prediction profile $l$ that occurs infinitely often, we have

$$\Pr\left( \left| f_T^l - \frac{2l^j - 1}{2n} \right| > \frac{1}{2n} + 2\varepsilon \right) > 1 - \varepsilon/2. \tag{16}$$

whenever $T$ is sufficiently large. Pick such a period $T$ for which, in addition, forecasting according to the true distribution $P$ ensures that one will be cross-calibrated within $\varepsilon$ with probability at least $1-\varepsilon/2$. After $T$ periods, with probability at least $1-\varepsilon$, the cross-calibration score of the true expert is higher by at least $\varepsilon$ than that of the

false expert. In particular, by choosing fine enough intervals for the cross-calibration test, the finite horizon approximation to the cross-calibration test can only be passed by predictions that are close to the true distribution. Since the true expert is likely to be selected, we also have that the predictions are close to the actual process.

Although the cross-calibration test succeeds, with high probability, in identifying the true expert in finite time (hence also approximately identifying the true probabilities), this time is not bounded and generally depends on the experts' strategies. The test used by Al-Najjar and Weinstein (2006) compares the ratio of probabilities that a true expert and a false expert predicted for the realization. They show that the test identifies predictions close to the true probabilities. Interestingly, it does so, with high probability, in all but a finite number of periods that is uniformly bounded independent of the experts' strategies.

We conclude this section with the following proposition for the single-expert calibration tests. It states that when using a pure strategy—following some distribution $P$—a potential expert can be calibrated on at most a category one set of realizations. Naturally, that category one set of realizations has $P$-probability one, as shown by Dawid (1982). This demonstrates that the calibration test is a particular example of a good test as defined by Dekel and Feinberg (2006). Combining this proposition with the main result of Foster and Vohra (1998), it follows immediately that there exists a good test that can be manipulated. See Olszewski and Sandroni (2006b), who were first to demonstrate the existence of such a test. We will use the following proposition in the proof of our main theorem below.

**Proposition 3** *For every $P$, the set of realizations at which $P$ is calibrated is a category one set. Hence, calibration is a good test.*

**Proof.** We will prove the result for the weak calibration test as defined by Kalai, Lehrer and Smorodinsky (1999), which requires calibration only for sequences of predictions that occur with positive density (a positive proportion of the time). The set of realizations on which $P$ passes the (standard) calibration test is a subset of the set of

realizations on which $P$ passes the weak calibration test since the standard test requires that additional conditions be satisfied. It suffices to prove the result for every partition of the form $[0,x],[x,1]$ for $x \in (0,1)$, since calibration on a partition $\{I_1,\ldots,I_n\}$ implies calibration on $\{I_1, I_2 \cup \cdots \cup I_n\}$.

We define the density of a sequence of time periods $T_1 < T_2 < ... < T_n < ...$ as

$$\limsup_{n\to\infty} \frac{n}{T_n}. \tag{17}$$

Note that the sequence does not have positive density if and only if the sequence $\{n/T_n\}_{n=1}^{\infty}$ converges to zero.

Let $P \in \Delta(\Omega)$ and $x \in (0,1)$ be given. Define the sets

$$S_{n,m,0} = \left\{ \omega \mid \sum_{t=0}^{m-1} \frac{\mathbb{1}_{p_t(\omega_t)\le x}}{m} \le 1/n \text{ or } \frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega_t)\le x}\omega_{t+1}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega_t)\le x}} \le x + 1/n \right\} \tag{18}$$

and

$$S_{n,m,1} = \left\{ \omega \mid \sum_{t=0}^{m-1} \frac{\mathbb{1}_{p_t(\omega_t)\ge x}}{m} \le 1/n \text{ or } \frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega_t)\ge x}\omega_{t+1}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega_t)\ge x}} \ge x - 1/n \right\} \tag{19}$$

and let

$$S_n^M = \bigcap_{m=M}^{\infty} \left( S_{n,m,0} \cap S_{n,m,1} \right). \tag{20}$$

Letting $N = \max\{10, 2/x, 2/(1-x)\}$, define the set

$$S = \bigcap_{n=N}^{\infty} \bigcup_{M=1}^{\infty} S_n^M. \tag{21}$$

Let $\omega$ be such that $P$ is weakly calibrated at $\omega$. In particular, for $I = [0,x]$, we either have

$$\limsup_{T\to\infty} \sum_{t=0}^{T-1} \frac{\mathbb{1}_{p_t(\omega_t)\in I}}{T} = 0 \tag{22}$$

or

$$\limsup_{T\to\infty}{}^{*} \frac{\sum_{t=0}^{T} \mathbb{1}_{p_t(\omega_t)\in I}\omega_{t+1}}{\sum_{t=0}^{T} \mathbb{1}_{p_t(\omega_t)\in I}} \le x, \tag{23}$$

where the notation $\liminf^*, \limsup^*$ refers to limits taken only over sequences with positive density (these are the relevant limits for weak calibration). We claim that for every $n \geq N$, $\omega \in S_{n,m,0}$ for all $m$ sufficiently large. If Equation (22) holds, then for every $n \geq N$ there exists some $M$ such that for all $m \geq M$,

$$\sum_{t=0}^{m-1} \frac{\mathbb{1}_{p_t(\omega_t) \leq x}}{m} \leq 1/n. \tag{24}$$

Hence for every $n \geq N$, there exists some $M$ such that $\omega \in S_{n,m,0}$ for all $m \geq M$. If, on the other hand, Equation (22) does not hold, then assume for contradiction that there exists some $n > N$ such that $\omega \notin S_{n,m,0}$ for an infinite sequence of values of $m$. Since (23) holds, this sequence cannot have positive density, which implies that for all large enough $m$ in this sequence, Inequality (24) holds, contradicting that $\omega$ does not belong to any of these sets.

The symmetric argument applied to the interval $I = [x, 1]$ demonstrates that for every $n \geq N$, $\omega \in S_{n,m,1}$ for all sufficiently large $m$. Combining these two results, we find that for every $n \geq N$, there exists some $M$ such that $\omega \in S_n^M$, and therefore $\omega \in S$. In addition, if $P$ is not weakly calibrated in either $[0, x]$ or $[x, 1]$ at $\omega$, then $\omega \notin S$, for there exists some $n > N$ and infinitely many $m$ with $\omega \notin S_{n,m,*}$. Hence for some $n$, we have $\omega \notin S_n^M$ for all $M$, which implies that $\omega \notin S$.

We need to show that $S$ is a category one set in $\Omega$. We will show that each $S_n^M$ is a closed set with empty interior. Since $S$ is a countable intersection of a countable union of such sets, it is a category one set. The set $S_n^M$ is an intersection of sets of the form $S_{n,m,l}$ with $l \in \{0, 1\}$, so it will be closed if all of the sets $S_{n,m,l}$ are closed. Without loss of generality, consider the case $l = 0$. For every $\omega \notin S_{n,m,0}$, we have

$$\sum_{t=0}^{m-1} \frac{\mathbb{1}_{p_t(\omega_t) \leq x}}{m} > 1/n \tag{25}$$

and

$$\frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega_t) \leq x} \omega_{t+1}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega_t) \leq x}} > x + 1/n. \tag{26}$$

Consider every $\omega'$ such that $\omega'_t = \omega_t$ for $t = 0, ..., m-1$. Since both conditions above depend only on the first $m$ coordinates of $\omega$, each such $\omega'$ is not a member of $S_{n,m,0}$. This collection of these $\omega'$ constitute a finite cylinder and hence comprise an open set. Therefore, every point outside $S_{n,m,0}$ has an open neighborhood outside this set and $S_{n,m,0}$ is closed.

Consider any point $\omega \in S_n^M$. Let $x \leq 1/2$. Define a sequence of realizations $(\omega(j))_{j=1,2,...}$ by

$$\omega(j)_t = \begin{cases} \omega_t & \text{if } t \leq j \\ 1 & \text{if } t > j \text{ and } p_{t-1}(\omega(j)_t | \omega(j)_1, \ldots, \omega(j)_{t-1}) < x \\ 0 & \text{if } t > j \text{ and } p_{t-1}(\omega(j)_t | \omega(j)_1, \ldots, \omega(j)_{t-1}) \geq x. \end{cases} \tag{27}$$

By definition, $\omega(j)$ agrees with $\omega$ in the first $j$ coordinates; hence the sequence $(\omega(j))_{j=1,2,...}$ converges to $\omega$. It suffices to show that $\omega(j) \notin S_n^M$. If the density of $P$ at $[x, 1]$ given $\omega(j)$ is at least $1/10$, then there exist infinitely many $m > M$ such that

$$\sum_{t=0}^{m-1} \frac{\mathbb{1}_{p_t(\omega(j)_t) \geq x}}{m} > 1/10 \geq 1/N \geq 1/n. \tag{28}$$

For $m$ large enough, we also have

$$\frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t) \geq x} \omega(j)_{t+1}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t) \geq x}} < x/2 = x - x/2 < x - 1/N \leq x - 1/n \tag{29}$$

since, for $t \geq j$, $\omega(j)_{t+1} = 0$ whenever $\mathbb{1}_{p_t(\omega(j)_t) \geq x} = 1$ and so the empirical distribution converges to zero. We conclude that if the density of $P$ at $[x, 1]$ for $\omega(j)$ is at least $1/10$ then $\omega(j) \notin S_n^M$.

If the density $\eta$ at $[x, 1]$ is less than $1/10$, then the density $\rho$ in $[0, x)$ must be at least $1 - \eta > 9/10$ since the sum of densities for the intervals $[0, x), [x, 1]$ must be at least 1. For infinitely many $m > M$, we have

$$\sum_{t=0}^{m-1} \frac{\mathbb{1}_{p_t(\omega(j)_t) \leq x}}{m} > 9/10 \geq 1/N \geq 1/n. \tag{30}$$

The empirical frequency when $[0, x]$ is predicted is given by

$$\frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)\leq x}\omega(j)_{t+1}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)\leq x}} = \frac{\sum_{t=0}^{m-1} (\mathbb{1}_{p_t(\omega(j)_t)=x}\omega(j)_{t+1} + \mathbb{1}_{p_t(\omega(j)_t)<x}\omega(j)_{t+1})}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)\leq x}}. \quad (31)$$

When $\mathbb{1}_{p_t(\omega(j)_t)=x}$ we have $\omega(j)_{t+1} = 0$, and when $\mathbb{1}_{p_t(\omega(j)_t)<x}$ we have $\omega(j)_{t+1} = 1$. Substituting these into Equation (31) gives

$$\frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)\leq x}\omega(j)_{t+1}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)\leq x}} = \frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)<x}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)\leq x}} \geq \frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)<x}}{m}. \quad (32)$$

By the definition of the density at $[0, x)$, there exist infinitely many $m$ such that

$$\frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)<x}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)\leq x}} \geq \frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)<x}}{m} \geq \rho - 1/10 > 1/2 + 2/10 > x + 1/N \geq x + 1/n, \quad (33)$$

indicating that $\omega(j) \notin S_n^M$, as required.

For the case where $x > 1/2$, define the sequence $\omega(j)$ as in (27) except with the inequality on the second line weak, and on the third line strong. Applying the symmetric argument with the roles of the intervals $[0, x]$ and $[x, 1]$ reversed gives the result. ■

## 4.2 No True Experts

When there are no true experts, a false expert who does not know the strategies of others remains unable to manipulate the test since he can at most manipulate against a category one set of *pure* strategies. In particular, there does not exist a strategy that succeeds in manipulating the test against all strategies of the other false experts. But what if the false experts know each other's strategies? Can they jointly manipulate the test? A single false expert who knows the strategies employed by other false experts can choose a manipulating strategy. This follows from the observation that fixing the others' strategies in the cross-calibration test is equivalent to the randomized calibration tests studied by Lehrer (2001) for the false expert who knows these strategies. Lehrer showed that all such tests are manipulable. To eliminate joint manipulation, we consider a stronger test—the strict cross-calibration test. This test requires that

predictions made infinitely often in adjacent, yet different, intervals do not simultaneously pass the test. Moving to the strict test introduces some Type I errors, as true distributions with conditional probabilities that converge to the edge of the predicted interval may fail the test. Olszewski and Sandroni (2006b) have also studied tests that reject some distributions out of hand. They show that by allowing some Type I errors, the tester can prevent a false expert from arbitrarily delaying rejection in a finite time approximation test.

We show that manipulation by at most one false expert is all that the strict cross-calibration allows even when the experts are informed of one another's strategies. Moreover, there is no equilibrium as long as each expert prefers to pass the test: if one expert passes with positive probability, then the other expert must be using a strategy for which the probability of passing the test is less than one. Hence the second expert could, by choosing a different strategy that passes almost surely, be strictly more likely to pass the test. But this latter strategy must leave the first expert using a strategy that passes the test with zero probability, and cannot be a best response. We conclude that the false experts cannot jointly manipulate the strict test.

We assume throughout that experts cannot correlate their randomized predictions. While they can condition on all past realized randomized predictions, they cannot use correlated strategies. With correlated strategies, multiple false experts can act as a single expert and joint manipulation would arise. Our last result (Proposition 4 below) considers the case where the experts are allowed to make precise predictions; that is, instead of predicting intervals, they provide exact conditional probabilities. It turns out that a sequential test of this sort can be manipulated simultaneously since the experts can exploit the richness of the allowed forecasts in the first period to signal a full strategy for all remaining periods.

**Theorem 1** *For any strategy profile* $\mu = (\mu^1, ..., \mu^M)$ *of* $M \geq 2$ *false experts, the set of realizations on which at least two false experts simultaneously pass the strict cross-calibration test with positive probability is a category one set in* $\Omega$.

**Proof.** Fix the realization $\omega$. For each $\eta \in (\frac{1}{2}, 1)$, define the (possibly empty) set

$$H_\eta = \{\text{finite histories } h \mid \Pr(\text{forecasts agree forever after } h | h, \omega) > \eta\}.$$

Note that histories include the realizations of previous forecasts. Let $\overline{H}_\eta$ denote the complement of $H_\eta$ in the set of all finite histories. The following lemma states that the probability that the forecasters are simultaneously strictly cross-calibrated without reaching any history in $H_\eta$ is zero.

**Lemma 2** *Fix $\eta \in (\frac{1}{2}, 1)$ and the realization $\omega$. If $\Pr(H_\eta) < 1$, then*

$$\Pr(i \textit{ and } j \textit{ are strictly cross-calibrated} \mid \overline{H}_\eta, \omega) = 0.$$

**Proof of Lemma 2.** We will assume throughout that $\omega$ is given and that all probabilities are conditional on $\omega$ and $\overline{H}_\eta$. Suppose that $\Pr(H_\eta) < 1$. Then there exists some $t_0$ such that $\Pr\left(\exists s \leq t_0 \text{ such that } I_s^i \neq I_s^j\right) \geq (1-\eta)/2$, since otherwise the empty history would lie in $H_\eta$, and we would have $\Pr(H_\eta) = 1$. By the same argument for every history $h_{t_0} \in \overline{H}_\eta$ of length $t_0$ that occurs with positive probability, there exists $t(h_{t_0}) > t_0$ such that for the history $h_{t_0}$, we have

$$\Pr\left(\exists s \in (t_0, t(h_{t_0})] \text{ such that } I_s^i \neq I_s^j | h_{t_0}\right) \geq (1-\eta)/2.$$

Since the number of histories of length $t_0$ is finite, by choosing $t_1 = \max_{h_{t_0}} t(h_{t_0})$, we have

$$\Pr\left(\exists s \in (t_0, t_1)] \text{ such that } I_s^i \neq I_s^j | h_{t_0}\right) \geq (1-\eta)/2.$$

Inductively, there is a finite $t_j$ such that

$$\Pr\left(\exists s \in (t_{j-1}, t_j)] \text{ such that } I_s^i \neq I_s^j | h_{t_{j-1}}\right) \geq (1-\eta)/2, \tag{34}$$

and this also holds for every history of length $t_{j-1}$ in $\overline{H}_\eta$ that occurs with positive

probability. Define the events

$$E_j = \exists s \in (t_{j-1}, t_j] \text{ such that } I_s^i \neq I_s^j. \tag{35}$$

We have that

$$\Pr\left(\neg E_{j+k} | \neg E_j \cap \neg E_{j+1} \cap \cdots \cap \neg E_{j+k-1}\right) \leq \eta/2 \tag{36}$$

since (34) holds for every history $h_{t_{j-1}} \in \overline{H}_\eta$ that has positive probability.

Considering the event that the forecasts do not disagree infinitely often (i.o.), we have that

$$\begin{aligned}\Pr(\text{ the complement of } E_j \text{ holds i.o.}) &= \Pr\left(\neg \bigcap_{n=1}^{\infty} \bigcup_{j \geq n} E_j\right) \\ &= \Pr\left(\bigcup_{n=1}^{\infty} \bigcap_{j \geq n} \neg E_j\right) \leq \sum_{n=1}^{\infty} \Pr\left(\bigcap_{j \geq n} \neg E_j\right).\end{aligned} \tag{37}$$

For every $n$, we have from (36) that

$$\begin{aligned}\Pr\left(\bigcap_{j \geq n} \neg E_j\right) &= \Pr(\neg E_n) \Pr(\neg E_{n+1} | \neg E_n) \cdots \Pr(\neg E_{n+k} | \neg E_n \cap E_{n+1} \cap \cdots \cap E_{n+k-1}) \cdots \\ &\leq \frac{\eta}{2}\frac{\eta}{2} \cdots \to 0, \end{aligned} \tag{38}$$

which, together with (37), implies that

$$\Pr(\text{ the complement of } E_j \text{ holds i.o.}) = 0.$$

Since forecasts that are strictly cross-calibrated cannot disagree infinitely often, the probability of strict cross-calibration must be 0. ■

Fix a realization $\omega$ on which the probability $\gamma$ that two forecasters simultaneously pass the strict cross-calibration test is positive. For each $\eta \in (\frac{1}{2}, 1)$, Lemma 2 implies

that there exists some $h_T \in H_\eta$ that occurs with positive probability and satisfies

$$\Pr(i \text{ and } j \text{ are strictly cross-calibrated} \mid h_T, \omega) \geq \gamma. \tag{39}$$

We will show that when $\eta$ is sufficiently large, following the history $h_T$, there is a particular path of forecasts that occurs with probability greater than $1-\gamma$ on which the players agree in every period. In particular, the forecasters both pass the strict cross-calibration test on this path.

For each finite history $h$, let $p(h) = (p_1(h), \ldots, p_n(h))$ and $q(h) = (q_1(h), \ldots, q_n(h))$ denote the mixed forecasts of the two players in the period immediately following $h$. For each $h$, there exists some $l(h) \in \{1, \ldots, n\}$ satisfying $p_{l(h)}(h) \geq \frac{1}{n}$.

We define a path of forecasts following the given history $h_T$, i.e. a unique path of realizations of the forecasts of the players given $\omega$ and the realization of forecasts $h_T$ after which both agree. We will show that this path occurs with high probability when $\eta$ is close to 1. For each $t > T$, recursively define the history $h_t = \big(h_T, \omega_T, I_{l(T)}, I_{l(T)}, \ldots, \omega_{t-1}, I_{l(t-1)}, I_{l(t-1)}\big)$, where each $l(\tau)$ satisfies $p_{l(\tau)}(h_{\tau-1}) \geq \frac{1}{n}$ (if there exists more than one such $l(\tau)$, then the choice among them is arbitrary). For each $l \in \{1, \ldots, n\}$ and $t \geq T$, let $\rho_l^t = p_l(h_t)q_l(h_t)$ denote the probability that both players forecast $I_l$ in the period following $h_t$.

**Lemma 3** *Conditional on the history $h_T$ of* (39) *occurring, the infinite history $\big(h_T, \omega_T, I_{l(T)}, I_{l(T)}, \omega_{T+1}, I_{l(T+1)}, I_{l(T+1)}, \ldots\big)$ recursively defined above occurs with probability greater than $n(\eta-1)+1$.*

**Proof of Lemma 3.** Once again all probabilities are conditional on $\omega$. We have

$$\Pr(\text{forecasts agree forever after } h_T | h_T) \leq \prod_{t \geq T} \rho_{l(t)}^t + \sum_{t \geq T} \left( \prod_{\tau=T}^{t-1} \rho_{l(\tau)}^\tau \right) \sum_{l' \neq l(t)} \rho_{l'}^t. \tag{40}$$

The first term on the right-hand side represents the probability of remaining on the specified path. The second term is an upper bound on the probability of leaving this path, but agreeing in every period nonetheless. This term captures, for each $t$, the

probability that the first deviation from the most likely path occurs in period $t$, and yet the forecasts agree in that period. We will show that

$$\sum_{t\geq T}\left(\prod_{\tau=T}^{t-1}\rho_{l(\tau)}^{\tau}\right)\sum_{l'\neq l(t)}\rho_{l'}^{t}<(n-1)(1-\eta), \tag{41}$$

which, together with (40) and the fact that $h_T\in H_\eta$, implies that

$$\prod_{t\geq T}\rho_{l(t)}^{t}<\eta-(n-1)(1-\eta)=n(\eta-1)+1,$$

as needed.

Note that

$$\Pr(\text{forecasts disagree in some period after } h_T|h_T)\geq\sum_{t\geq T}\left(\prod_{\tau=T}^{t-1}\rho_{l(\tau)}^{\tau}\right)\left(1-\sum_{l}\rho_{l}^{t}\right), \tag{42}$$

since the $t$ term on right-hand side of this inequality is the probability of remaining on the specified path until period $t$, at which time the players choose two different forecasts. Since $h_T\in H_\eta$, the probability of disagreement after $h_T$ is less than $1-\eta$, and hence (42) implies

$$\sum_{t\geq T}\left(\prod_{\tau=T}^{t-1}\rho_{l(\tau)}^{\tau}\right)\left(1-\sum_{l}\rho_{l}^{t}\right)<1-\eta. \tag{43}$$

We will show that, for all $t\geq T$,

$$\sum_{l'\neq l(t)}\rho_{l'}^{t}\leq(n-1)\left(1-\sum_{l}\rho_{l}^{t}\right), \tag{44}$$

from which, together with (43), the desired inequality (41) follows immediately.

Since $p_{l(t)}(h_t)\geq\frac{1}{n}$, we have

$$\begin{aligned}\left(1-p_{l(t)}(h_t)\right)\left(1-q_{l(t)}(h_t)\right)&\leq\left(1-\frac{1}{n}\right)\left(1-q_{l(t)}(h_t)\right)\\&\leq(n-1)\left(p_{l(t)}(h_t)\left(1-q_{l(t)}(h_t)\right)+q_{l(t)}(h_t)\left(1-p_{l(t)}(h_t)\right)\right).\end{aligned}$$

Rearranging terms gives

$$n\left(1-p_{l(t)}(h_t)\right)\left(1-q_{l(t)}(h_t)\right) \leq (n-1)\left(1-p_{l(t)}(h_t)q_{l(t)}(h_t)\right).$$

Note that

$$\sum_{l\neq l(t)} p_{l(t)}(h_t)q_{l(t)}(h_t) \leq \left(1-p_{l(t)}(h_t)\right)\left(1-q_{l(t)}(h_t)\right)$$

since the left-hand side represents the probability (following $h_t$) that both players announce the *same* forecast other than $I_{l(t)}$, whereas the right-hand side represents the probability that neither announces $I_{l(t)}$. Combining the last two inequalities gives

$$n\sum_{l\neq l(t)} p_{l(t)}(h_t)q_{l(t)}(h_t) \leq (n-1)\left(1-p_{l(t)}(h_t)q_{l(t)}(h_t)\right),$$

which is simply a rearrangement of (44), as needed. ■

For $\gamma > 0$, consider the set of realizations $\tilde{\Omega}(\gamma) \subset \Omega$ on which the experts are simultaneously strictly cross-calibrated with probability at least $\gamma$. The set of realizations on which the experts are simultaneously cross-calibrated with positive probability is a countable union $\bigcup_n \tilde{\Omega}(\gamma_n)$ of these sets, where $\gamma_n \to 0$. Thus it suffices to show that $\tilde{\Omega}(\gamma)$ is a category one set for each $\gamma > 0$.

Fixing an arbitrary $\gamma > 0$, we will write $\tilde{\Omega}$ in place of $\tilde{\Omega}(\gamma)$. By Proposition 3, a countable collection of pure strategies in the classic calibration test can only pass on a category one set of realizations. Hence the proof of the theorem is complete if we can show that there exists such a collection of strategies out of which, for each realization in $\tilde{\Omega}$, at least one is calibrated.

As noted above, by choosing $\eta$ sufficiently close to 1, Lemmas 2 and 3 together imply that, for each $\omega \in \tilde{\Omega}$, there exists some history $h_T$ after which both players are strictly cross-calibrated if they follow the path of forecasts $I_{l(t)}$. Fix such a history for each $\omega \in \tilde{\Omega}$, and for each finite history $h_T$, let $\tilde{\Omega}(h_T) \subset \tilde{\Omega}$ denote the realizations associated with $h_T$ in this way.

Having fixed the history $h_T$ for each realization, let $l_\omega(t)$ denote the forecast $l(t)$

of Lemma 3 given $\omega$. To each history $h_T$ for which $\tilde{\Omega}(h_T)$ is nonempty, associate the pure strategy $p^{h_T}$ defined by

$$p^{h_T}(h_t) = \begin{cases} I(\omega_t) & \text{if } t \leq T \\ I_{l_\omega(t)} & \text{if } t > T \text{ and } h_t \text{ agrees with } \omega \in \tilde{\Omega}(h_T) \\ I(\frac{1}{2}) & \text{otherwise,} \end{cases} \tag{45}$$

where, for $t \leq T$, $\omega_t$ denotes the $t$-coordinate of the realization in $h_T$ (recall that $h_T$ represents the realization of $\omega$ together with the realized forecasts). As long as $\eta > 1 - \frac{1}{2n}$, each $l(t)$ in Lemma 3 occurs with probability greater than $\frac{1}{2}$, and is therefore unique. Moreover, by construction, $l_\omega(t)$ depends only on the past history at time $t$, not on the future realization of $\omega$. Therefore, the strategy $p^{h_T}$ is well-defined. Since the false experts are strictly cross-calibrated at $\omega \in \tilde{\Omega}(h_T)$ if they follow the forecasts $I_{l_\omega(t)}$ following $h_T$, $p^{h_T}$ is calibrated at $\omega$.

We have shown that $\tilde{\Omega}$ is a subset of the realizations for which one of the countable collections of pure strategies $p^{h_T}$ is calibrated. Since, by Proposition 3, each pure strategy is calibrated on a category one set of realizations, the set $\tilde{\Omega}$ is itself a category one set, and the proof of the theorem is complete. ■

We conclude by showing that allowing false experts to announce precise predictions rather than limiting them to choosing an interval leads to joint manipulability. This result follows from the observation that a single real number can be used to encode each distribution $P$. Thus whenever a single expert has a randomized strategy that can manipulate a test, he can encode each realization of this mixed strategy within his first prediction, allowing all other false experts to perfectly correlate their randomization by following the distribution corresponding to this first prediction. Correlating in this way guarantees that all false experts simultaneously manipulate the test.

**Proposition 4** *Consider a cross-calibration test where the experts report real-valued predictions (the test classifies these predictions according to the intervals to which they belong). There exists a strategy profile such that the false experts can manipulate the*

*test jointly as long as no true expert is present.*

**Proof.** A single false expert can manipulate the calibration test according to Lehrer (2001). Hence there is a mixed strategy $\mu \in \Delta(\Delta(\Omega))$ for the false expert, such that for every $\omega$, the realized pure forecasting strategy $P \in \Delta(\Omega)$ almost surely (with respect to $\mu$) passes the calibration test on $\omega$. Consider such a mixed strategy $\mu$ for a single potential expert. Order the finite histories $\omega_t$ of possible realizations, for example according to $\emptyset, 0, 1, 00, 01, 10, 11, 000, 001, \ldots$. Each distribution $P$ over $\Omega$ is defined by a sequence of real numbers $r_1, r_2, \ldots, r_k, \ldots$, which denote for each finite history the probability that a 1 occurs in the following period. A mixed strategy is a distribution over $\Delta(\Omega)$, which induces a distribution over these sequences of real numbers.

We define a mapping $g : [0,1]^\infty \to [0,1]$ and show that it is measurable and one-to-one. Consider $(r_1, r_2, \ldots, r_k, \ldots) \in [0,1]^\infty$. Let $r_k = (r_k^1, r_k^2, ...)$ be such that $r_k^i \in \{0, 1, \ldots, 9\}$ is the $i$th digit in the decimal expansion of $r_k$ (let 1 be represented by $0.999\ldots$). Define $g(r_1, r_2, \ldots, r_k, \ldots) = (g^1, g^2, \ldots)$ to be the real number with expansion $(g^1, g^2, \ldots)$ satisfying $g^2 = r_1^1, g^4 = r_1^2, \ldots$, and in general, $g^{(p_k)^n} = r_k^n$ where $p_k$ denotes the $k$th prime number and $(p_k)^n$ is its $n$th power. When $l$ is not a prime power, set $g^l = 0$.

The mapping $g$ is one-to-one since two different sequences are always mapped to numbers that differ in some coordinate. The measurability of $g$ follows from the observation that for every coordinate, a measurable set of reals is mapped to a measurable set of reals. Note that the function $g$ is in fact continuous except for the sets of sequences that have a coordinate where the expansion is not unique (this exceptional set is countable).

Now consider the realization of the single expert's manipulative strategy $\mu$, and let the strategies of multiple experts be determined as follows: One of the experts randomizes according to $\mu$, giving rise to a realization $P$. Before the first period, this expert predicts the real number $g(r_1, r_2, \ldots, r_k, \ldots)$, where $r_1, r_2, \ldots, r_k, \ldots$ is the sequence of probabilities corresponding to $P$. All other experts make the prediction 0.5 before the first period, and after that follow the pure strategy $P$ as defined by

$g(r_1, r_2, \ldots, r_k, \ldots)$ (when a deviation occurs and the first expert does not choose a number in the range of $g$ all experts play some arbitrary strategy). Since $g$ is one-to-one and measurable, these strategies are well-defined. All experts will make identical predictions after the first period, and hence they are cross-calibrated if and only if one of them is calibrated in the single expert test. The result follows by the choice of $\mu$. ■

# References


[1] Al-Najjar, N. I., and Weinstein J. (2006) "Comparative Testing of Experts." *Mimeo.*

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

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

[4] Dekel, E., and Feinberg, Y. (2006) "Non-Bayesian Testing of a Stochastic Prediction." *The Review of Economic Studies* **72** (4), 893-906.

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

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

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

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

[9] Olszewski, W. and Sandroni, A. (2006a) "Counterfactual Predictions." *Mimeo.*

[10] Olszewski, W. and Sandroni, A. (2006b) "Strategic Manipulation of Empirical Tests." *Mimeo.*

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

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

[13] Vovk V., and Shafer, G. (2005) "Good randomized sequential probability forecasting is always possible." J. R. Stat. Soc. Ser. B Stat. Methodol. **67** , no. 5, 747-763.