Testing Multiple Forecasters*

Yossi Feinberg† Colin Stewart‡

Stanford University University of Toronto

January 2008

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.

*We wish to thank Nabil Al-Najjar, Brendan Beare, Dean Foster, Sergiu Hart, Stephen Morris, Wojciech Olszewski, Larry Samuelson, Alvaro Sandroni, Jakub Steiner, Jonathan Weinstein, three anonymous referees, and seminar participants at Austin, Ben-Gurion University, Bogota, Essex, Hebrew University, Michigan, Rice, San Diego, Stanford and the Technion for helpful comments and suggestions. The first author gratefully acknowledges the support of the NSF grant IIS-0205633 and the hospitality of the Institute for Advanced Studies at the Hebrew University.

†email: yossi@gsb.stanford.edu

‡email: colin.stewart@utoronto.ca

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 sequential test revealing their types by pitting their predictions against one another.

An intuitive sequential test asks that the expert's predictions be calibrated, i.e. that the empirical frequency 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 by Kalai, Lehrer, and Smorodinsky (1999); Fudenberg and Levine (1999); Lehrer (2001); Sandroni, Smorodinsky, and Vohra (2003); Sandroni (2003); and Vovk and Shafer (2005). See also Fortnow and Vohra (2006) and Chang and Lyuu (2007), who study testing from a computational perspective. Recently, Olszewski and Sandroni (2007) and Shmaya (2007) obtained the strong result that *all* sequential tests of a single potential expert can be manipulated.

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; we call it 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 30%. Conditional on these predictions, the empirical frequency cannot be both 40% *and* 30%. 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 respect to this test. This feature plays a central role in the greater power of the cross-calibration test relative to the classic calibration test. In independent work, Al-Najjar and Weinstein (2007) consider a different test which compares the likelihoods of predictions made by a false and a true expert. We discuss their work in detail in Section 5.

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. On the other hand, a false expert is guaranteed to fail the test for most distributions P when a true expert is present. More precisely, we show that, for every (mixed) forecasting strategy a false expert may use, he fails the cross-calibration test with P-probability one on all but a category one set of true distributions P.

Even when there is no true expert, a strict version of the cross-calibration test possesses some power to fail false experts. The strict test requires the empirical frequencies to lie within the predicted intervals and not on their boundaries. We show that, except on a small set of realizations, the probability that at least two potential experts pass this test simultaneously is zero.

Finally, we show that the realizations on which a pure forecasting strategy P is calibrated (in the classic calibration test of a single forecaster) form a category one

set. Hence calibration is a good test with no Type I errors and small Type II error in terms of category (see Dekel and Feinberg (2006)). In particular, cross-calibration is also a good test since it too has no Type I error and small Type II error; the set of realizations on which a forecaster is cross-calibrated is a subset of those on which he is calibrated when tested in isolation.

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 with the set of distributions over any finite set S, and the intervals I_l are replaced with a cover of the set of distributions by sufficiently small closed convex subsets.

At the beginning of each period $t = 0, 1, \ldots$, all forecasters (or experts) $j \in \{1, \ldots, M\}$ 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 forecasters at the end of each period. A (mixed or behavior) 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 Ω 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 func-

tions

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

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 $\omega_t \in \{0,1\}$, 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.

We consider two types of 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. 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 without loss of generality that the lower interval is chosen. Note that, although the expert uses a strategy that follows the true distribution P, he provides only the conditional probabilities for the realized history. Thus it is not necessary that the true expert know P *ex ante*; it suffices for him to know the conditional probabilities once a history is realized.

False experts have no knowledge of Nature's strategy. 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.

To minimize notation, we provide proofs of results for just two 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 μ and P.

3 With True Experts

We begin by exploring the outcome of the cross-calibration test when at least one of the potential experts is indeed a true expert. We first observe that no matter what others

do, every true expert is guaranteed to pass the cross-calibration test with P-probability one. Hence, the cross-calibration test has no Type I error.

Proposition 1 *For every distribution P governing the stochastic process, any potential expert who predicts according to a model that follows P passes 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 first forecaster passes the cross-calibration test with (P, μ)-probability one:*

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

Proof. The proof requires a minor modification of Dawid (1982), and is omitted. See Feinberg and Stewart (2007) for a complete proof. ■

We now turn to the case of false experts being tested in the presence of a true expert. The failure probability is determined according to P and the strategies employed by false experts. We would like to see the false experts fail with probability one according to the true distribution P, except perhaps for a small set of true distributions (since the false expert may happen to make predictions very close to the correct ones). We show that not only is a false expert unable to manipulate cross-calibration, but that for any strategy he might use, he is guaranteed to fail on *most* true distributions. This result contrasts sharply with the negative results in the single-expert case. Calibration-type tests of a single expert can be manipulated in the sense that a false expert can pass the test with μ-probability one on *every* realization, and hence for every true distribution P.

The notion for large and small sets of distributions we employ is that of category one—a countable union of nowhere dense sets—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.

The basic intuition for this result is as follows. In order for two potential experts to pass the cross-calibration test on the same realization, their forecasts must be close in all but finitely many periods. Since the true expert passes with probability one, the false expert must announce forecasts close to the truth in all but finitely many periods in order to pass the test. In order for this to occur with positive probability, there must be some finite history after which the false expert makes forecasts close to the truth in every period with *high* probability. For a given forecasting strategy μ, only a small set of distributions give conditional probabilities close to forecasts generated by μ with high probability after some history.

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

Proof. We first prove the following lemma.

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

$$\Pr(\mu \textit{ is close to } P \textit{ in every period following } h_T^\eta \mid h_T^\eta) \geq 1 - \eta. \tag{5}$$

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, (5) does not hold for the empty history. We can therefore find a finite time t_0 such that

$$\Pr\left(\exists s \leq t_0 \text{ such that } \mu \text{ is not close to } P \text{ at period } s\right) \geq \eta/2. \tag{6}$$

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

$$\Pr\left(\exists s \in (t_{j-1}, t_j] \text{ such that } \mu \text{ is not close to } P \text{ at period } s|h_{t_{j-1}}\right) \geq \eta/2 \tag{7}$$

for every $h_{t_{j-1}}$ that occurs with positive probability.

Define the events

$$F_j = \left\{\exists s \in (t_{j-1}, t_j] \text{ such that } \mu \text{ is not close to } P \text{ at period } s\right\}. \tag{8}$$

The event that the forecasts are close from some period onwards is the complement of the forecasts being not close infinitely often. To be cross-calibrated the experts must predict close intervals from some point onwards. Hence

$$\begin{aligned}\Pr\left(\text{The experts are cross-calibrated}\right) &\leq \Pr\left(\neg \bigcap_{n=1}^{\infty} \bigcup_{j \geq n} F_j\right) \\ &= \Pr\left(\bigcup_{n=1}^{\infty} \bigcap_{j \geq n} \neg F_j\right) \leq \sum_{n=1}^{\infty} \Pr\left(\bigcap_{j \geq n} \neg F_j\right).\end{aligned} \tag{9}$$

We now show that

$$\Pr\left(\bigcap_{j \geq n} \neg F_j\right) = 0 \tag{10}$$

for every n. If $\Pr\left(\neg F_n \cap \neg F_{n+1} \cap ... \cap \neg F_{n+k-1}\right) = 0$ for some $k > 0$, then (10) holds trivially. Otherwise, since (7) holds for every history $h_{t_{j-1}}$ that occurs with positive probability, it also holds when conditioned on any positive probability collection of

histories $h_{t_{j-1}}$. In particular, we have

$$\begin{aligned}\Pr\left(\neg F_{n+k}|\neg F_n \cap \neg F_{n+1} \cap ... \cap \neg F_{n+k-1}\right) &= 1 - \Pr\left(F_{n+k}|\neg F_n \cap \neg F_{n+1} \cap ... \cap \neg F_{n+k-1}\right) \\ &\leq 1 - \eta/2.\end{aligned}$$

Therefore, for every n, we have

$$\begin{aligned}\Pr\left(\bigcap_{j\geq n} \neg F_j\right) &= \Pr\left(\neg F_n\right) \Pr\left(\neg F_{n+1}|\neg F_n\right) \cdots \Pr\left(\neg F_{n+k}|\neg F_n \cap F_{n+1} \cap ... \cap F_{n+k-1}\right) \cdots \\ &\leq \left(1 - \frac{\eta}{2}\right)\left(1 - \frac{\eta}{2}\right) \cdots = 0. \quad (11)\end{aligned}$$

From (9) and (10) we have that the experts are cross-calibrated with probability zero—a contradiction, as required. ■

Fix $\eta < \frac{1}{2}$. By Lemma 1, if P and μ are cross-calibrated with positive probability, then P must satisfy (5) for at least one of the countable collection of finite histories. It suffices to show that the set of distributions that satisfy (5) 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

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

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 ε to $\Omega(h)$ *and* for which h has the property of (5) (given η and μ). The set of distributions against which μ 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 ε to $\Omega(h)$, or h does not satisfy (5) (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 $\eta' > \eta$ such that

$$\lim_{t \to \infty} \Pr\left(\mu \text{ is close to } P \text{ in every period from } T+1 \text{ to } t | h\right) < 1 - \eta'.$$

Hence there exists some period τ such that

$$\Pr\left(\mu \text{ is close to } P \text{ in every period from } T+1 \text{ to } \tau | h\right) < 1 - \eta'. \tag{12}$$

Each distribution P' gives rise for each t to an induced distribution P'_t over finite histories of realizations $(\omega_0, \ldots, \omega_t)$. Let $H_\tau = \{h_t \text{ consistent with } h \text{ for } t = T+1, \ldots, \tau\}$.

Assume first that P satisfies

$$p_t\left(h_t\right) \notin \left\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, 1\right\} \tag{13}$$

for every $h_t \in H_\tau$ that occurs with positive P-probability.[1] Consider the set

$$U_\delta = \left\{P' \mid |P'_\tau(E) - P_\tau(E)| < \delta \quad \text{for all events } E \text{ determined by time } \tau\right\}.$$

Note that U_δ is open for each δ since it is defined by a strict inequality condition on a collection of (open) cylinders.

By (13), we can find $\delta > 0$ sufficiently small such that, in any period following any $h_t \in H_\tau$, μ is close to $P' \in U_\delta$ if and only if μ is close to P. For such δ, (12) implies

[1] Recall that $p_t(h_t)$ denotes the conditional P-probability that $\omega_t = 1$ following h_t.

that

$$\Pr_P\left(\mu \text{ is close to } P' \text{ in every period from } T+1 \text{ to } \tau|h\right) < 1-\eta'.$$

By the definition of U_δ, this last inequality implies that

$$\Pr_{P'}\left(\mu \text{ is close to } P' \text{ in every period from } T+1 \text{ to } \tau|h\right) < 1-\eta$$

when δ is sufficiently small. This guarantees that U_δ is disjoint from $S(h,\varepsilon)$, as needed.

If (13) does not hold, then there exists some finite history h_t occurring with positive P-probability such that $\Pr_P(\omega_t = 1|h_t) \in \{0, 1/n, \ldots, 1\}$. We will show that the set of distributions P having this property is category one, and hence adding it to the union of sets $S(h,\varepsilon^n)$ does not affect the claim. Since the set of finite histories is countable and the set $\{0, 1/n, \ldots, 1\}$ is finite, it suffices to show that, given any finite history h_t, any $\pi \in [0,1]$, and any $\varepsilon > 0$, the set

$$R(h_t,\varepsilon,\pi) = \{P|p_t(h_t) = \pi \text{ and } P_t(h_t) \geq \varepsilon\}$$

is closed with empty interior.

First we show that $R(h_t,\varepsilon,\pi)$ is closed. Let $S_\varepsilon = \{P|P_t(h_t) \geq \varepsilon\}$. Note that S_ε is closed. Let A denote the event that h_t occurs, and B the event that $\omega_t = 1$. Consider the function $f_{B|A} : \Omega \longrightarrow \mathbb{R}$ defined by

$$f_{B|A}(\omega) = \begin{cases} \pi & \text{if } \omega \notin A \\ 1 & \text{if } \omega \in A \cap B \\ 0 & \text{otherwise.} \end{cases}$$

Note that, since A and B are clopen, $f_{B|A}$ is continuous. Hence the set of distributions $S_\pi = \left\{P| \int f_{B|A} dP = \pi\right\}$ is closed in the weak* topology. For each $P \in S_\pi$, either $P(A) = 0$ or $P(B|A) = \pi$. Therefore, we have $R(h_t,\varepsilon,\pi) = S_\pi \cap S_\varepsilon$, proving that $R(h_t,\varepsilon,\pi)$ is closed since S_π and S_ε are.

Next we show that $R(h_t, \varepsilon, \pi)$ has empty interior. Fix any $\delta \in (0, \pi)$. Given any $P \in R(h_t, \varepsilon, \pi)$, let $(P^n)_{n=1}^{\infty}$ be the sequence of distributions with conditional probabilities

$$p_\tau^n(\omega_0, \ldots, \omega_{\tau-1}) = \begin{cases} \pi - \delta^n & \text{if } (\omega_0, \ldots, \omega_{\tau-1}) = h_t \\ p(\omega_0, \ldots, \omega_{\tau-1}) & \text{otherwise.} \end{cases}$$

The sequence $(P^n)_{n=1}^{\infty}$ converges to P and lies outside of the set $R(h_t, \varepsilon, \pi)$, as needed.

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 $\eta < \frac{1}{2}$, μ 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 distribution against which he almost surely fails. Moreover, the set of such distributions is large in the sense that its complement in $\Delta(\Omega)$ is a category one set in the weak* topology. To obtain some intuition for the meaning of a category one set of measures in $\Delta(\Omega)$, recall that such a set is a countable union of nowhere dense sets. Any nowhere dense set $S \subset \Delta(\Omega)$ has the following property: for any finitely determined event E and any $P \in S$, there exist measures outside S that agree with P on E, and for every measure P' outside S, there exists a finite event on which P' differs from *all* measures in S. In particular, the probabilities assigned to finitely determined events can never rule out distributions outside the nowhere dense set.

We note that our proof implies that a false expert will fail cross-calibration in finite time with high probability. In fact, such finite approximation results hold for all limit results in this paper, much like the finite approximation results in Dekel and Feinberg (2006) and Olszewski and Sandroni (2007). In particular, fix a mixed strategy μ and a true distribution P such that the probability that μ 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

$$\Pr\left(\left|f_T^l - \frac{2l^j - 1}{2n}\right| > \frac{1}{2n} + 2\varepsilon\right) > 1 - \varepsilon/2 \tag{14}$$

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 ε 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 ε 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.

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).[2] Combining this proposition with the main result of Foster and Vohra (1998), it follows immediately that calibration is a good test that can be manipulated. See Olszewski and Sandroni (2006), who were the first to demonstrate the existence of a good manipulable test. We will use the following proposition in the proof of Proposition 4 below.

[2]A test is called *good* if it has no Type I errors and small Type II errors, i.e., for each true distribution P, a forecaster using P passes the test with P-probability one, and for any distribution Q, the set of true distributions for which predicting according to Q passes with positive probability is category one.

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 predictions that occur with positive density. The density of a sequence periods $T_1 < T_2 < \cdots < T_n < \cdots$ is defined as $\limsup_{n\to\infty} \frac{n}{T_n}$. 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. It suffices to prove the result for all subintervals of the form $[0, x]$, $[x, 1]$ for $x \in (0, 1)$, since calibration on the subintervals $\{I_1, \ldots, I_n\}$ implies calibration on $\{I_1, I_2 \cup \cdots \cup I_n\}$.

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

$$S_{n,m,0} = \left\{ \omega \mid \sum_{t=0}^{m-1} \frac{\mathbb{1}_{p_t(\omega_t)\leq x}}{m} \leq 1/n \text{ or } \frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega_t)\leq x}\omega_{t+1}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega_t)\leq x}} \leq x + 1/n \right\} \tag{15}$$

and

$$S_{n,m,1} = \left\{ \omega \mid \sum_{t=0}^{m-1} \frac{\mathbb{1}_{p_t(\omega_t)\geq x}}{m} \leq 1/n \text{ or } \frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega_t)\geq x}\omega_{t+1}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega_t)\geq x}} \geq x - 1/n \right\}, \tag{16}$$

and let

$$S_n^M = \bigcap_{m=M}^{\infty} \left(S_{n,m,0} \cap S_{n,m,1} \right). \tag{17}$$

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

$$S = \bigcap_{n=N}^{\infty} \bigcup_{M=1}^{\infty} S_n^M. \tag{18}$$

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

$$\limsup_{T\to\infty} \sum_{t=0}^{T-1} \frac{\mathbb{1}_{p_t(\omega_t)\in I}}{T} = 0 \tag{19}$$

or

$$\limsup_{T\to\infty}{}^{*} \frac{\sum_{t=0}^{T} \mathbb{1}_{p_t(\omega_t)\in I}\omega_{t+1}}{\sum_{t=0}^{T} \mathbb{1}_{p_t(\omega_t)\in I}} \leq x, \tag{20}$$

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 (19) holds, then for every $n \geq N$, there exists some M such that for all $m \geq M$,

$$\sum_{t=0}^{m-1} \frac{\mathbb{1}_{p_t(\omega_t) \leq x}}{m} \leq 1/n. \tag{21}$$

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 (19) 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 (20) holds, this sequence cannot have positive density, which implies that for all large enough m in this sequence, Inequality (21) holds, contradicting that ω 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 ω, 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 Ω. 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

$$\sum_{t=0}^{m-1} \frac{\mathbb{1}_{p_t(\omega_t) \leq x}}{m} > 1/n \tag{22}$$

and

$$\frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega_t) \leq x} \omega_{t+1}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega_t) \leq x}} > x + 1/n. \tag{23}$$

Consider every ω' such that $\omega'_t = \omega_t$ for $t = 0, ..., m-1$. Since both conditions above depend only on the first m coordinates of ω, each such ω' is not a member of $S_{n,m,0}$. The collection of these ω' 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

$$\omega(j)_t = \begin{cases} \omega_t & \text{if } t \leq j \\ 1 & \text{if } t > j \text{ and } p_{t-1}(\omega(j)_t | \omega(j)_1, \ldots, \omega(j)_{t-1}) < x \\ 0 & \text{if } t > j \text{ and } p_{t-1}(\omega(j)_t | \omega(j)_1, \ldots, \omega(j)_{t-1}) \geq x. \end{cases} \tag{24}$$

By definition, $\omega(j)$ agrees with ω in the first j coordinates; hence the sequence $(\omega(j))_{j=1,2,...}$ converges to ω. 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

$$\sum_{t=0}^{m-1} \frac{\mathbb{1}_{p_t(\omega(j)_t) \geq x}}{m} > 1/10 \geq 1/N \geq 1/n. \tag{25}$$

For m large enough, we also have

$$\frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t) \geq x} \omega(j)_{t+1}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t) \geq x}} < x/2 = x - x/2 < x - 1/N \leq x - 1/n \tag{26}$$

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 frequency 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 η at $[x, 1]$ is less than $1/10$, then the density ρ 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

$$\sum_{t=0}^{m-1} \frac{\mathbb{1}_{p_t(\omega(j)_t) \leq x}}{m} > 9/10 \geq 1/N \geq 1/n. \tag{27}$$

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

$$\frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)\leq x}\omega(j)_{t+1}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)\leq x}} = \frac{\sum_{t=0}^{m-1} (\mathbb{1}_{p_t(\omega(j)_t)=x}\omega(j)_{t+1} + \mathbb{1}_{p_t(\omega(j)_t)<x}\omega(j)_{t+1})}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)\leq x}}. \quad (28)$$

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 (28) gives

$$\frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)\leq x}\omega(j)_{t+1}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)\leq x}} = \frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)<x}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)\leq x}} \geq \frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)<x}}{m}. \quad (29)$$

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

$$\frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)<x}}{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)\leq x}} \geq \frac{\sum_{t=0}^{m-1} \mathbb{1}_{p_t(\omega(j)_t)<x}}{m} \geq \rho - 1/10 > 1/2 + 2/10 > x + 1/N \geq x + 1/n, \quad (30)$$

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

For the case where $x > 1/2$, define the sequence $\omega(j)$ as in (24) except with the inequality on the second line weak, and the inequality 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 Without True Experts

When there is no true expert, a false expert cannot manipulate the test for all strategies of the other forecasters since, by Proposition 2, he can manipulate against only a category one set of *pure* strategies. On the other hand, for any strategy profile of the other forecasters, there exists a strategy that will almost surely pass the test on every realization. This follows from the observation that, once the opponents' strategies are fixed, the cross-calibration test becomes equivalent to a randomized calibration test (of a single forecaster) of the class studied by Lehrer (2001). Lehrer showed that such tests are manipulable. This result extends to any sequential test with no Type I error according to footnote 8 in Olszewski and Sandroni (2007), see also Shmaya (2007).

In the absence of a true expert, therefore, such a test can at best be guaranteed to fail *all but one* false expert. The question remains whether *multiple* false experts can manipulate the test simultaneously.

We show that multiple false experts cannot jointly manipulate a stronger version of our test—the strict cross-calibration test. Whatever forecasting strategies these false experts use, no two of them can pass the strict test with positive probability except on a small set of realizations.

If the conditional empirical frequency 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 $\{[0, 1/n), [1/n, 2/n), ..., [(n-1)/n, 1]\}$. Thus we modify the inequality in (3) to a strict inequality (on one side of the interval) when needed, to reflect that the empirical frequency must converge to the appropriate interval.

A true expert may fail the strict cross-calibration test. For example, if the true distribution gives rise to independent probabilities $1/n - 1/e^t$ in each period t, then the empirical frequency converges to $1/n$ from below with probability one. Olszewski and Sandroni (2006) 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 assume throughout that experts cannot correlate their randomized predictions. While they are allowed to condition on all past realizations of randomized predictions, they cannot use correlated strategies. Otherwise, false experts could act as one and manipulate the test.[3]

Proposition 4 *For any strategy profile $\mu = (\mu^1, ..., \mu^M)$ of $M \geq 2$ (false) experts, the set of realizations on which at least two potential experts simultaneously pass the strict cross-calibration test with positive probability is a category one set in Ω.*

[3]See Feinberg and Stewart (2007) for an explicit proof of correlated manipulation with real-valued predictions.

Proof. Fix the realization ω. 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_η 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_η is zero.

Lemma 2 *Fix $\eta \in (0,1)$ and the realization ω. 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. The proof is essentially the same as for Lemma 1, and is therefore omitted. There are only two significant differences. First, the condition $I_s^i \neq I_s^j$ replaces the condition that μ is not close to P at period s. Second, probabilities are now with respect to forecasting strategies μ^i and μ^j given a fixed realization ω, instead of being with respect to the forecasting strategy μ and the true distribution P. ■

Fix a realization ω on which the probability γ 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

$$\Pr(i \text{ and } j \text{ are strictly cross-calibrated} \mid h_T, \omega) \geq \gamma. \tag{31}$$

We will show that when η 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 forecasters 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 forecasters 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 experts given ω and the realization of forecasts h_T after which both agree. We will show that this path occurs with high probability when η is close to 1. For each $t > T$, recursively define the history $h_t = \left(h_T, \omega_T, I_{l(T)}, I_{l(T)}, \ldots, \omega_{t-1}, I_{l(t-1)}, I_{l(t-1)}\right)$, 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 forecasters forecast I_l in the period following h_t.

Lemma 3 *Let* $\overline{h} = \left(h_T, \omega_T, I_{l(T)}, I_{l(T)}, \omega_{T+1}, I_{l(T+1)}, I_{l(T+1)}, \ldots\right)$*, as defined above for* h_T *satisfying* (31)*. We have*

$$\Pr\left(\overline{h}|h_T, \omega\right) > n(\eta - 1) + 1.$$

Proof of Lemma 3. Once again all probabilities are conditional on ω. Note first that

$$\begin{aligned}\sum_{t \geq T} \left(\prod_{\tau=T}^{t-1} \rho_{l(\tau)}^{\tau}\right)\left(1 - \sum_l \rho_l^t\right) &\leq \Pr(\text{forecasts disagree in some period after } h_T|h_T) \\ &< 1 - \eta, \end{aligned} \tag{32}$$

since the t term on left-hand side is the probability of remaining on the specified path until period t, at which time the forecasters choose two different forecasts. The second inequality follows since $h_T \in H_\eta$, i.e., the probability of disagreement after h_T is less than $1 - \eta$.

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

$$\begin{aligned}\left(1 - p_{l(t)}(h_t)\right)\left(1 - q_{l(t)}(h_t)\right) &\leq \left(1 - \frac{1}{n}\right)\left(1 - q_{l(t)}(h_t)\right) \\ &\leq \left(1 - \frac{1}{n}\right)\left(1 - p_{l(t)}(h_t)q_{l(t)}(h_t)\right). \end{aligned} \tag{33}$$

Note that

$$\sum_{l \neq l(t)} p_l(h_t) q_l(h_t) \leq \left(1 - p_{l(t)}(h_t)\right) \left(1 - q_{l(t)}(h_t)\right) \tag{34}$$

since the left-hand side represents the probability (following h_t) that both forecasters announce the *same* forecast other than $I_{l(t)}$, whereas the right-hand side represents the probability that neither announces $I_{l(t)}$. From (33) and (34), we get

$$n \sum_{l \neq l(t)} p_l(h_t) q_l(h_t) \leq (n-1) \left(1 - p_{l(t)}(h_t) q_{l(t)}(h_t)\right), \tag{35}$$

which rearranged yields

$$\sum_{l \neq l(t)} \rho_l^t \leq (n-1) \left(1 - \sum_l \rho_l^t\right) \tag{36}$$

for every $t \geq T$.

Inequalities (32) and (36) imply

$$\sum_{t \geq T} \left(\prod_{\tau=T}^{t-1} \rho_{l(\tau)}^{\tau} \right) \sum_{l \neq l(t)} \rho_l^t < (n-1)(1-\eta). \tag{37}$$

We also have that

$$\Pr(\text{forecasts agree forever after } h_T | h_T) \leq \prod_{t \geq T} \rho_{l(t)}^t + \sum_{t \geq T} \left(\prod_{\tau=T}^{t-1} \rho_{l(\tau)}^{\tau} \right) \sum_{l' \neq l(t)} \rho_{l'}^t, \tag{38}$$

since the first term on the right-hand side represents the probability of remaining on the specified path and the second term is an upper bound on the probability of leaving this path, but nonetheless agreeing in every period. 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. Since $h_T \in H_\eta$, the left-hand side of (38) is greater than η, and hence combining (37) and (38) gives

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

which completes the proof. ■

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 γ. 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 proposition is complete if we can show that there exists such a countable collection of strategies out of which, for each realization in $\tilde{\Omega}$, at least one is calibrated.

As noted above, by choosing η 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 forecasters 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 ω. To each history h_T for which $\tilde{\Omega}(h_T)$ is nonempty, associate the pure strategy p^{h_T} defined by

$$p^{h_T}(h_t) = \begin{cases} I(\omega_t) & \text{if } t \leq T \\ I_{l_\omega(t)} & \text{if } t > T \text{ and } h_t \text{ agrees with } \omega \in \tilde{\Omega}(h_T) \\ I(\frac{1}{2}) & \text{otherwise,} \end{cases} \tag{39}$$

where, for $t \leq T$, ω_t denotes the t-coordinate of the realization in h_T (recall that h_T represents the realization of ω 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 ω. 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 ω.

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 proposition is complete. ■

Proposition 4 indicates not only that multiple experts cannot simultaneously manipulate, but that all but one are *guaranteed* to fail on all but a category one set of realizations. The proof exploits the fact that, in order for two forecasters to pass the strict test simultaneously, they must announce identical predictions in all but finitely many periods. Lemmas 2 and 3 show that, for this to occur with positive probability (given some realization ω), there must be some history after which both forecasters are likely to predict according to a particular path of forecasts. But then there is a pure strategy that both forecasters' predictions are likely to follow, in which case they can pass the test only if this pure strategy passes the classic calibration test at ω. By Proposition 3, this can happen only on a category one set of realizations.

Since a category one set of realizations has positive probability according to at most a category one set of distributions (Dekel and Feinberg (2006)), Proposition 4 implies:

Corollary 1 *Fix any strategy profile μ of $M \geq 2$ (false) experts. For all but a category one set of true distributions P, at least $M-1$ experts will fail the strict cross-calibration test with (μ, P)-probability one.*

5 Related Literature and Discussion

As noted in the Introduction, the literature on testing forecasters has focused primarily on negative results for *sequential* tests of a single potential expert. The principle underlying these results is a Minmax, or separation, type theorem. If each prediction according to a true distribution must get a passing score, then under appropriate conditions, there is a randomized prediction strategy—a mixed strategy which induces a behavior strategy—that passes the test no matter what Nature does (i.e. for *every*

realization).

These negative 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 shown by Dekel and Feinberg (2006).[4] While a potential expert has the same set of strategies when facing a sequential and an ex ante test, the set of ex ante *tests* is larger. In particular, sequential tests can use only one sequence of realized predictions, which endows them with a continuity property enabling manipulation.

With multiple experts, new possibilities arise. When a true expert is known to be present, the question becomes *which* of the experts is the true expert, rather than *whether* a potential expert is a true one. In independent work, Al-Najjar and Weinstein (2007) consider this case, and propose a test which compares the likelihoods of predictions made by a false and a true expert. Their test selects the expert who is more likely to know the truth based on updating an equal prior probability. Technically, their test compares the product of the probabilities assigned to the realized outcome by each forecaster. Al-Najjar and Weinstein work mostly with the finite time version of this comparative test. They elegantly show that a false expert cannot manipulate their test: for every forecasting strategy the false expert might use, there exists a true distribution for which he is likely to lose. This is similar to the finite approximation of the cross-calibration test discussed in Section 3 above. However, there are a number of differences between the results.

Al-Najjar and Weinstein's results provide a uniform bound on the time required to approximately identify the true probabilities; no matter what strategy the false expert uses, he is unlikely to pass the test unless he announces probabilities close to the truth in all but a fixed number of periods. For cross-calibration, on the other hand, the time required is finite, but may not be uniformly bounded.[5] Another difference between our

[4]See also Olszewski and Sandroni (2006) for a stronger result.

[5]All non-manipulation results for infinite tests extend to finite approximations in this way quite generally (cf. Dekel and Feinberg (2006) and Olszewski and Sandroni (2007)). Moving from finite to infinite tests, however, is more difficult since it requires some consistency across the finite distributions on which manipulation fails.

results and theirs is that we provide a bound on the set of distributions on which a false expert passes—at most a category one set. Hence, non-manipulability is assured for most true distributions, not just one. Finally, because it is comparative, Al-Najjar and Weinstein's test has no power to prevent joint manipulation when all experts choose their forecasts strategically.

The topological notion of category one characterizes the extent of manipulability both in the space of distributions when a true expert is present, and in the space of realizations when one is not. Since category one is not commonly used in the economics literature, we describe some of its properties in our setting. By Proposition 4 in Feinberg and Dekel (2006), any infinite test for which passing occurs only on a category one set has a finite approximation for which passing requires assigning high probability to a nowhere dense set. In particular, these finite approximations rule out manipulation except when the false expert happens to assign positive probability to distributions concentrated on a nowhere dense set of realizations. Category one is the currently best known bound for manipulation, due to Olszewski and Sandroni (2006).[6] We find it attractive since the set of all distributions does not allow for a uniform measure, and hence there is no natural alternative candidate for a measure-theoretic notion of smallness. Furthermore, a nowhere dense set of distributions is also small in the sense that identifying whether a distribution belongs to this set requires an infinite amount of data: for every finite time, the probabilities of events defined up to that period never rule out distributions outside the set. Perhaps the most interesting phenomenon is that this same notion appears as a bound on manipulation both with and without true experts present.

Following the literature on testing forecasters, we have focused on a non-Bayesian setting. Hence we address "worst-case scenario" types of questions. For testing without a true expert, these questions include the following:

- Can one false expert manipulate the test when given the strategies of the other

[6]Feinberg and Dekel (2006) showed that their test cannot be manipulated on a category two set, a weaker notion than the complement of a category one set.

forecasters?

- Can multiple false experts jointly manipulate the test when they can correlate their forecasts?
- Can multiple false experts jointly manipulate the test when they cannot correlate their forecasts?

Answering the first and second questions is straightforward. A single expert, when given the strategies of others, can manipulate the test. Similarly, if the experts can correlate their forecasts, then they can jointly manipulate the test. Our main result in this setting answers the third question: multiple experts cannot jointly manipulate the strict cross-calibration test. Testing multiple forecasters also suggests new avenues of research. Even if false experts *can* use correlated forecasts to manipulate the test, it is possible that they may not *want* to. This issue raises the question of whether the tester could provide incentives for experts to counter collusive correlation. Answering this question would require an explicit formulation of the experts' incentives in order to apply game-theoretic tools in this non-Bayesian setting.

References

[1] Al-Najjar, N. I., and Weinstein J. (2007) "Comparative Testing of Experts." *Mimeo.*

[2] Chang, C.-L., and Lyuu, Y.-D. (2007) "Efficient Testing of Forecasts." *Lecture Notes in Computer Science* **4598**, 285–295.

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

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

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

[6] Feinberg, Y., and Stewart, C. (2007) "Testing Multiple Forecasters." *Stanford University Graduate School of Business Research Paper No.* 1957.

[7] Fortnow, L., and Vohra, R. V. (2007) "The Complexity of Forecast Testing." *Mimeo.*

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

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

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

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

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

[13] Olszewski, W. and Sandroni, A. (2007) "Future-Independent Tests." *Mimeo.*

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

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

[16] Shmaya, E. (2007) "Many inspections are manipulable." *Mimeo.*

[17] 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.