"Calibeating": Beating Forecasters at Their Own Game*

Dean P. Foster† Sergiu Hart‡

April 15, 2026

Addendum:
Calibeating Is Stronger Than the Stronger Expert

We show here that calibeating is a stronger notion than the so-called "stronger expert."[1]

Let $C = \Delta(A)$; as we have seen at the end of Section 2 (see (2)), the refinement score is the minimal Brier score over all relabelings of the bins; i.e.,

$$\mathcal{R}_t = \min_{\phi} \mathcal{B}_t^{\phi(\mathbf{c})},$$

where the minimum is taken over all functions $\phi : \Delta(A) \to \Delta(A)$ (from current labels to new labels), and we write $\mathcal{B}_t^{\phi(\mathbf{c})}$ for the Brier score where the sequence $\mathbf{c}$ is replaced by $\phi(\mathbf{c}) = (\phi(c_t))_{t\geq 1}$.

Let $\mathbf{b}^1, \ldots, \mathbf{b}^N$ be N forecasting sequences, where for each $1 \leq n \leq N$ we have $\mathbf{b}^n = (b_t^n)_{t\geq 1}$ and there is a finite set B^n such that $b_t^n \in B^n$ for all $t \geq 1$.

Therefore, we have:

- $\mathbf{c}$ *multi-calibeats* $\mathbf{b}^1, \ldots, \mathbf{b}^N$ if

$$\mathcal{B}_t^{\mathbf{c}} \leq \min_{1\leq n\leq N} \min_{\phi^n} \mathcal{B}_t^{\phi^n(\mathbf{b}^n)} + o(1),$$

*Addendum to the published journal version (*Theoretical Economics*, 2023) and to `arXiv` version 2 (2022). Written in November 2024; updated in April 2026.

†Department of Statistics, Wharton, University of Pennsylvania, Philadelphia, and Amazon, New York. *e-mail*: `dean@foster.net` *web page*: `http://deanfoster.net`

‡Institute of Mathematics, Department of Economics, and Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem. *e-mail*: `hart@huji.ac.il` *web page*: `http://www.ma.huji.ac.il/hart`

[1]Expands on the reply given to a question asked at a lecture at the Workshop on Learning in Games, Toulouse, July 2024.

as[2] $t \to \infty$, where for each n the minimum is taken over all functions $\phi^n : B^n \to \Delta(A)$; and

- $\mathbf{c}$ *calibeats the joint* of $\mathbf{b}^1, \ldots, \mathbf{b}^N$ if[3]

$$\mathcal{B}_t^{\mathbf{c}} \leq \min_{\phi} \mathcal{B}_t^{\phi(\mathbf{b}^1,\ldots,\mathbf{b}^N)} + o(1)$$

as $t \to \infty$, where the minimum is taken over all functions $\phi : \Pi_{n=1}^N B^n \to \Delta(A)$.

By comparison, when all forecasts are probability distributions on A, i.e., all the sets B^n are subsets of $\Delta(A)$, Foster (1991) defines:[4]

- $\mathbf{c}$ is *as strong as* $\mathbf{b}^1, \ldots, \mathbf{b}^N$ if

$$\mathcal{B}_t^{\mathbf{c}} \leq \min_{1 \leq n \leq N} \mathcal{B}_t^{\mathbf{b}^n} + o(1),$$

as $t \to \infty$; and

- $\mathbf{c}$ is *as strong as the convex hull* of $\mathbf{b}^1, \ldots, \mathbf{b}^N$ if

$$\mathcal{B}_t^{\mathbf{c}} \leq \min_{w} \mathcal{B}_t^{w^1\mathbf{b}^1+\ldots+w^N\mathbf{b}^N} + o(1),$$

as $t \to \infty$, where the minimum is taken over all $w = (w^1, \ldots, w^N) \geq 0$ with $\sum_{n=1}^N w^n = 1$ (i.e., over all convex combinations of $\mathbf{b}^1, \ldots, \mathbf{b}^N$).

Thus, multi-calibeating is stronger than being as strong as, and calibeating the joint is stronger than being as strong as the convex hull. Whereas calibeating the joint, our way to achieve multi-calibeating, takes into account *all* bin relabelings, stronger-expert concepts do not go beyond linear-combination relabelings. Therefore, as claimed, *calibeating yields stronger notions than the "stronger expert."*

References

Foster, D. P. (1991), "Prediction in the Worst Case," *The Annals of Statistics* 19, 1084–1090.

Foster, D. P. and S. Hart (2023), "'Calibeating': Beating Forecasters at Their Own Game," *Theoretical Economics* 18, 1441–1474.

- Full version (2022): `http://arxiv.org/abs/2209.04892v2`

[2] For simplicity we ignore throughout the uniformity requirement with respect to the sequences $\mathbf{a}, \mathbf{b}^1, \ldots, \mathbf{b}^N$.

[3] Our way of achieving multi-calibeating is by calibeating the joint $\mathbf{b}^1 \times \ldots \times \mathbf{b}^N$; see Theorem 7.

[4] In the expansive literature on experts, these notions are referred to as "prediction with no regret"; see Appendix A.9 in the full paper for the parallel results with the logarithmic scoring rule.