THE AMERICAN STATISTICIAN
2019, VOL. 00, NO. 0, 1–6: General
https://doi.org/10.1080/00031305.2019.1604430

Taylor & Francis
Taylor & Francis Group

A Prediction Tournament Paradox

David J. Aldous

Department of Statistics, U.C. Berkeley, Berkeley, CA

ABSTRACT
In a prediction tournament, contestants "forecast" by asserting a numerical probability for each of (say) 100 future real-world events. The scoring system is designed so that (regardless of the unknown true probabilities) more accurate forecasters will likely score better. This is true for one-on-one comparisons between contestants. But consider a realistic-size tournament with many contestants, with a range of accuracies. It may seem self-evident that the winner will likely be one of the most accurate forecasters. But, in the setting where the range extends to very accurate forecasters, simulations show this is mathematically false, within a somewhat plausible model. Even outside that setting the winner is less likely than intuition suggests to be one of the handful of best forecasters. Though implicit in recent technical papers, this paradox has apparently not been explicitly pointed out before, though is easily explained. It perhaps has implications for the ongoing IARPA-sponsored research programs involving forecasting.

ARTICLE HISTORY
Received May 2018
Revised March 2019

KEYWORDS
Bradley–Terry model; Brier score; Competition; Forecasting; Probability assessment

1. Introduction

In mathematical terms, a prediction tournament consists of a collection of n questions of the form "state a probability for a specified real-world event happening before a specified date." In actual tournaments one can update probabilities as time passes, but for simplicity we consider only a single probability prediction for each question, and only binary outcomes. Scoring is by squared error[1]: if you state probability q then on that question

$$\text{score} = (1-q)^2 \text{ if event happens; score} = q^2 \text{ if not.}$$

Your tournament score is the sum of scores on each question. As in golf one seeks a *low* score. Also as in golf, in a *tournament* all contestants address the same questions; it is not a single-elimination tournament as in tennis. The use of squared-error scoring is designed so that (under your own belief) the expectation of your score is minimized by stating your actual probability belief. So in the long run it is best to be "honest" in that way.

General nonmathematical background to the topic is best found in the persuasive essay by Tetlock, Mellers, and Scoblic (2017). Study of results of prediction tournaments in recent years has led to the popular book by Tetlock and Gardner (2015) and substantial academic literature—for instance Mellers et al. (2014) has 147 Google Scholar citations. That literature involves serious statistical analysis, but is focused on the psychology of individual and team-based decision making and the effectiveness of training methods. This article addresses a conceptual point—how effective are tournaments at identifying the best forecasters?—via simulation and only quite elementary mathematics.

Here is the basic algebra. With unknown true probabilities (p_i), if you announce probabilities (q_i) then (see (1)) the true expectation of your score equals

$$\sum_i p_i(1-p_i) + \sum_i (q_i - p_i)^2.$$

The first term is the same for all contestants, so if S and $\hat{S}$ are the tournament scores for you and another contestant, then

$$n^{-1/2}(\mathbb{E}S - \mathbb{E}\hat{S}) = \sigma^2 - \hat{\sigma}^2$$

where

$$\sigma := \sqrt{n^{-1}\sum_i (q_i - p_i)^2}$$

is your RMS error in predicting probabilities and $\hat{\sigma}$ is the other contestant's RMS error. Thus by looking at differences in scores one can, in the long run, estimate relative abilities at prediction, as measured by RMS error of predicted probabilities.

To re-emphasize, when we talk about prediction ability we mean the ability to estimate *probabilities* accurately; we are not talking about predicting Yes/No outcomes and counting the number of successes, which is an extremely inefficient procedure for comparing prediction ability.

1.1. The Elementary Mathematics

Let us quickly write down the relevant elementary mathematics. Write X for your score on a question when the true probability is p and you predict q

$$\mathbb{P}(X = (1-q)^2) = p, \quad \mathbb{P}(X = q^2) = 1-p.$$
$$\mathbb{E}X = p(1-q)^2 + (1-p)q^2 = p(1-p) + (q-p)^2.$$

CONTACT David J. Aldous aldous@stat.berkeley.edu Department of Statistics, U.C. Berkeley, Berkeley, CA 94720-3860.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/TAS.
[1] In fact tournaments use *Brier score*, which is just 2× the squared error, with modifications for multiple-choice questions.

So writing S for your "tournament score" when the true probabilities of the n events are $(p_i, 1 \le i \le n)$ and you predict $(q_i, 1 \le i \le n)$,

$$\mathbb{E}S = \sum_i p_i(1-p_i) + n\sigma^2, \qquad (1)$$

where

$$\sigma^2 := n^{-1} \sum_i (q_i - p_i)^2$$

is your mean-squared error (MSE) in assessing the probabilities. Let us spell out some of the implications of this simple formula.

- The first term in (1) is the same for all contestants: one could call it the contribution from "irreducible randomness."
- The formula shows that a convenient way to measure forecast accuracy is via σ, the root-mean-square (RMS) error of a candidate's forecasts.
- The actual score S is random

$$S = \sum_i p_i(1-p_i) + \sigma^2 + \text{(chance variation)}, \qquad (2)$$

where the "chance variation" has expectation zero. Given the scores S and $\hat{S}$ for you and another contestant, one could attempt a formal test of significance of the hypothesis $\sigma < \hat{\sigma}$ that you are a more accurate forecaster. But making a valid test is quite complicated, because the "chance variations" are highly correlated.

1.2. But One-on-One Comparisons May Be Misleading

In the model and parameters we will describe in Section 2, a contestant in a 100-question tournament who is 5% more accurate than another (i.e., RMS prediction errors 10% vs. 15%, or 20% vs. 25%) will have around a 75% chance to score better (and around 90% chance if 10% more accurate). This is unremarkable; it is just like the well-known Bradley–Terry style models (see, e.g., Hunter 2004) for sports, where the probability A beats B is a specified function of the difference in strengths. In the sports setting it seems self-evident that in any reasonable league season or tournament play, the overall winner is likely to be one of the strongest teams. The purpose of this article is to observe, in the next section, that (within a simple model) this "self-evident" feature is just plain false for prediction tournaments. So in this respect, prediction tournaments are fundamentally different from sports contests.

Let us call this the *prediction tournament paradox*. Once observed, the explanation will be quite simple. Possible implications for real-world prediction tournaments will be discussed in Section 3.

2. Who Wins the Tournament?

Simulations in this section use the following model for a tournament.

> *Default tournament model*. There are 100 questions, with true probabilities 0.05, 0.15, 0.25, …, 0.95, each appearing 10 times.

This number of questions roughly matches the real tournaments we are aware of.

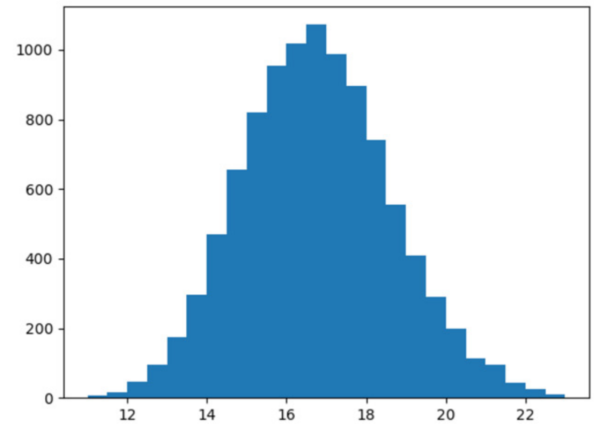

Figure 1. Chance variation in tournament score.

Table 1. Chance of more accurate forecaster beating less accurate forecaster in 100-question tournament.

		RMS error σ (less accurate)					
		0.05	0.1	0.15	0.2	0.25	0.3
	0	0.73	0.87	0.95	0.99	1.00	1.00
RMS	0.05		0.77	0.92	0.97	0.99	1.00
error σ	0.1			0.78	0.92	0.97	0.99
(more	0.15				0.76	0.92	0.97
(accurate)	0.2					0.76	0.91
	0.25						0.73

2.1. Intrinsic Variability

In this model, for a player who always predicted the true probabilities their mean score would be 16.75. But there is noticeable random variation between realizations of the tournament events, illustrated in Figure 1 histogram.

The variability in Figure 1 can be regarded (very roughly) as the "luck" in the 3-part decomposition (2).

2.2. Comparing Two Contestants

We do not expect contestants to predict exactly the true probabilities, so to understand a real tournament we need to model inaccuracy of predictions. This is conceptually challenging. The basic formula (1) shows that it is the MSE σ^2 in forecasting which affects score, so we parametrize "inaccuracy" by the RMS error σ. Amongst many possible models, we take what is perhaps the simplest.

> *Simple model for predictions by contestant with RMS error σ*. When the true probability is p, the contestant predicts $p \pm \sigma$, each with equal probability (independent for different questions, and truncated to $[0, 1]$).

Table 1 shows the probability that, in this model tournament, a more accurate forecaster gets a better score than a less accurate forecaster. The simulation results here correspond well to intuition; indeed the probability depends, roughly, on the difference in RMS errors.

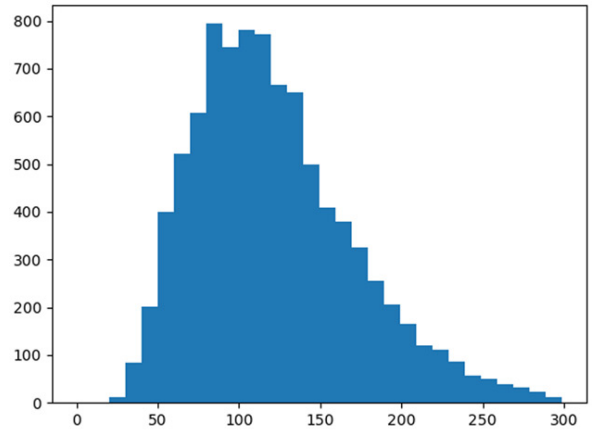

Figure 2. Rank of tournament winner, 300 contestants, error parameters $0 < \sigma < 0.3$.

2.3. Rank of Tournament Winner

We now consider a tournament with 300 contestants, keeping the model above for questions and forecasting accuracy. If all contestants had equal forecasting ability then each would be equally likely to be the winner. Modeling variability of accuracy amongst the field of contestants is also difficult, and again we take a simple model.

> *Simple model for variability of accuracy amongst contestants.* Abilities (measured by RMS error σ) range evenly across an interval, which we arbitrarily take to have length 0.3.

In pseudo-Python code

```
n_c = 300 # number of contestants
n_q = 100 # number of events
for i = 0 to n_q - 1:  p(i) = 0.05 + 0.1 floor(i/10) # True
probability of event i
      B(i) = Bernoulli(p(i)):  # outcome of event i
# [σ_0, σ_0 + 0.3] range of RMS errors
    for different contestants
for j = 0 to n_c - 1:  σ(j) = σ_0 + 0.3j/n_c:  # RMS error of
contestant j
      for i = 0 to n_q - 1:  q(i,j) = p(i) ± σ(j) #prediction
of contestant j
      for event i, with random ±
            score(i,j) = (B(i) - q(i,j))**2 #squared error,
contestant j event i
      score(j) = sum_i score(i,j) # total score
      of contestant j.
```

With 300 contestants, the top-ranked ability is little different from that of the second- or third-ranked, so the chance of the top-ranked contestant winning will not be large in absolute terms. But common sense and Table 1 results suggest the winner will be one of the relatively top-ranked contestants; as in any sport, the probability of being the tournament winner should decrease with rank of ability.

Figure 2 shows the results of the first simulation we did, taking the interval of RMS error parameters to be $[0, 0.3]$.

So here the winner is relatively most likely to be around the 100th most accurate of the 300 contestants, and the top-ranked contestants never win. This is in striking contrast to intuition—a paradox, in that sense. Indeed one might well suspect an error in coding the simulations. However, if we shift the assumed interval of σ successively to $[0.05, 0.35]$, $[0.1, 0.4]$, and $[0.15, 0.45]$ then (see Figure 3) we do soon see the intuitive "winning probability decreases with rank" property, but still the winners are not as strongly concentrated among the very best forecasters as one might have guessed.

2.4. First Explanation of the Paradox

Once observed, the original paradox is easy to explain in words. In the specific setting of Figure 2, the handful of top-rated contestants are making almost exactly the same predictions and therefore getting almost exactly the same score—as if there were just one such contestant. But looking at contestants with σ around 0.1 they are making slightly different predictions, on average scoring less well; but by chance, for some contestants, most of the predictions will vary in the direction of the outcome that actually occurred, and so these contestants will get a better score by pure luck. As a physical analogy, imagine contestants who each shoot successively at 100 different red targets. But there is an invisible-to-contestants blue target randomly displaced from each red target, and they are scored by the average distance between the shot and the blue target. The skillful contestants whose shots land close to the red targets will all get roughly the same score. Less skillful contestants will typically get lower scores, but some will by chance have more errors in the directions toward the blue targets and therefore, by luck, get a better score. This is an instance of a mean-variance trade-off. See Section 3 for further discussion.

One might wonder whether the specific implementation of RMS error σ as "predict $p \pm \sigma$ with equal probability" had any effect. Simulations with the alternative "predict a random variable uniform in $[p - \sigma\sqrt{3}, p + \sigma\sqrt{3}]$" implementation are shown in Figure 4, and the effect is even stronger in that model.

One might also wonder if this behavior is special to the winner, but we see a similar effect if we look at the ranks of contestants whose scores are in the top 10—see Figure 5.

Are these results merely artifacts of the specific model? Suppose instead we took a "one-sided" model of inaccuracy, say in the sense that contestants systematically over-estimate probabilities. Then there would be a comparatively large chance that the top-ranked contestant is the winner, from the case that more event outcomes than expected turned out to be "no." What about a model in which half the contestants systematically over-estimate probabilities and the other half systematically under-estimate? The results are shown in Figure 6. As before the inaccuracy parameter σ of different contestants varies evenly over $[0, 0.3]$, but now the prediction model is

> for half the contestants, when the true probability is p, the contestant predicts a random value uniform in $[p, p + \sigma\sqrt{3}]$ (independent for different questions, and truncated to $[0, 1]$); for the other half of the contestants,when the true

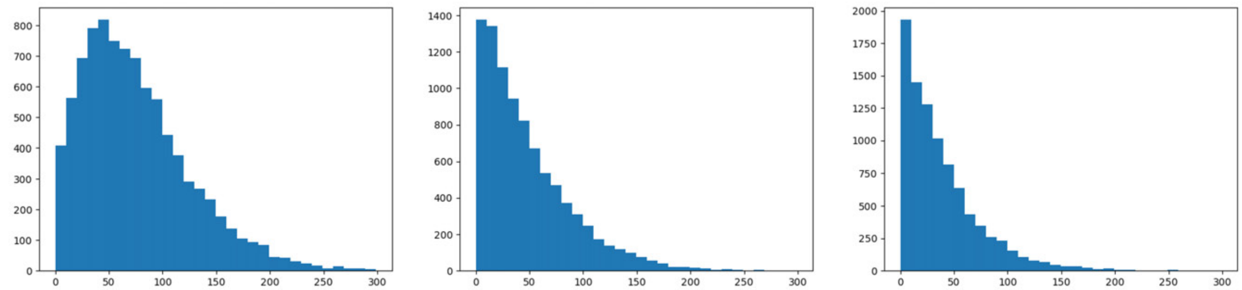

Figure 3. Rank of tournament winner, 300 contestants, error parameters $0.05 < \sigma < 0.35$ (left), $0.1 < \sigma < 0.4$ (center), and $0.15 < \sigma < 0.45$ (right).

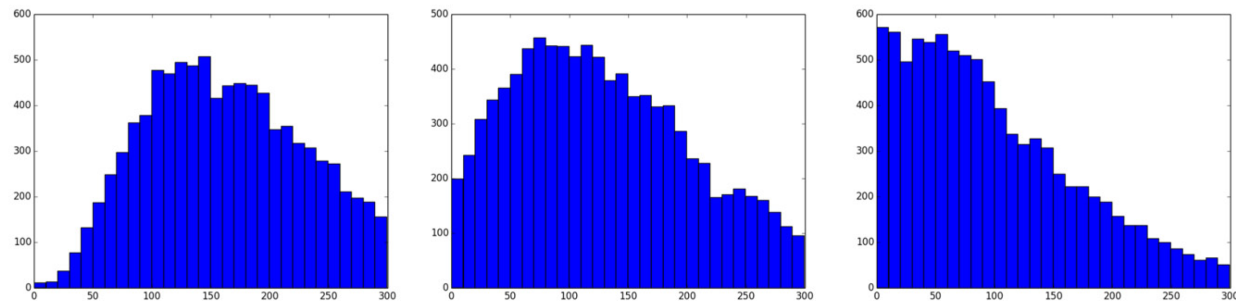

Figure 4. Alternate model: rank of tournament winner, 300 contestants, error parameters $0.05 < \sigma < 0.35$ (left), $0.1 < \sigma < 0.4$ (center), and $0.15 < \sigma < 0.45$ (right).

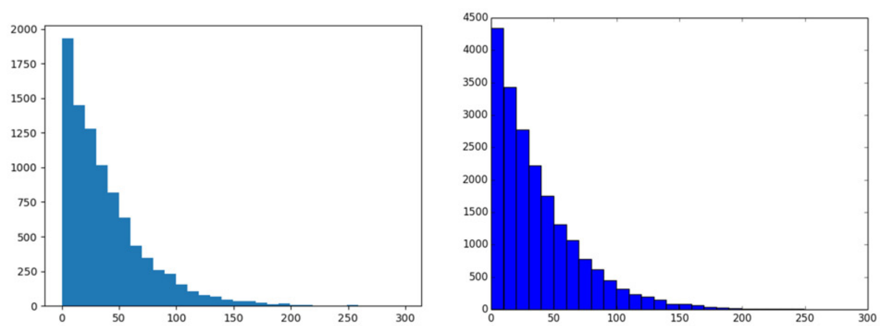

Figure 5. Ranks of tournament winner (left) and of top 10 finishers (right), $0.15 < \sigma < 0.45$.

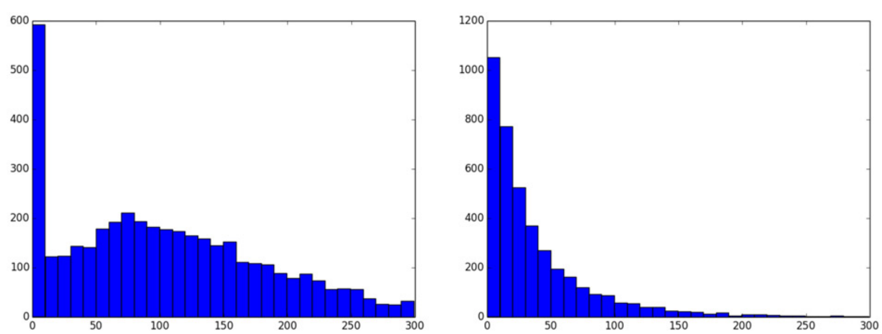

Figure 6. Rank of tournament winner, $0 < \sigma < 0.3$ (left) and $0.15 < \sigma < 0.45$ (right) for systematic over- or under-estimation.

probability is p, the contestant predicts a ran.om value uniform in $[p-\sigma\sqrt{3}, p]$ (independent for different questions, and truncated to $[0, 1]$);

Here, in the case $0 < \sigma < 0.3$ with accurate forecasters, we see a combination of the effects noted above. A near-top-rated contestant will likely win when the pattern of eventoutcomes is relatively close to balanced (events of probability p happen a proportion p of the time), but as in the previous one-sided case some of the biased contestants will, by luck, do better when outcomes are unbalanced. In the case $0.15 < \sigma < 0.45$ of all inaccurate forecasters the pattern of winner ranks is similar to the original model (Figure 3).

3. Discussion

One subtle point not discussed earlier is that our measure σ of accuracy is implicitly a long-term average. By modeling predictions as random, each contestant in our model has an empirical RMS prediction error in a finite tournament. In the size of tournaments simulated here, for contestants with equal σ the correlation between score and RMS empirical prediction error is small: the "noise" of event outcomes is overwhelming. One might hope that differences in long-term accuracy would show up in reasonable size tournaments, but part of our "paradox" is the observation that 100-question tournaments are not sufficient.

3.1. Explanations of the Paradox

To elaborate our earlier mean-variance-tradeoff explanation of the paradox with some numbers, consider a 100-question tournament in which the true probabilities are all 0.5. So a perfectly accurate forecaster will score exactly 25.0. Now consider a contestant who predicts 0.4 or 0.6 randomly on each question. Their score is random with expectation 26 and standard deviation (SD) 0.98, so have around 15% chance to beat the perfect forecaster. If instead predictions were 0.3 or 0.7, the expectation and SD become 29 and 1.83. Moreover, as a special feature of the "all true probabilities are 0.5" setting, different contestants' scores are independent. In our simulated setting of 300 contestants with RMS prediction errors ranging from 0 to 0.3. some scores will by chance be around 3 SDs below expectation, and by this back-of-an-envelope argument we expect a winning score around 23 and we will not be surprised if this comes from the 100th or 200th best forecaster.

Our simulations used the more plausible *default probability model* with varying true probabilities, and here there is a specific complicated dependence structure for the scores of different contestants, not amenable to convincing back-of-an-envelope calculations or convincing asymptotic approximations or human-interpretable algebra, which is why we have relied on simulations.

Experts in statistical methodology might readily think of their own explanations of or analogies for the paradox. For instance it is partly analogous to standard multiple comparison settings, though the specific dependence structure arising in this "estimating probabilities" context is quite different from the usual contexts of multiple comparisons for experimental or observational data. See Hung and Fithian (2019) for recent work on similar *rank verification* questions within the usual context.

In a typical sports setting, the winner of a tournament is indeed relatively more likely to be one of the best teams. So it is important to realize how our prediction tournament setting is conceptually very different from the more familiar setting of a contest in which each contestant earns points (directly reflecting skill, as in a basketball shot) in each of 100 rounds and the winner is the contestant with the most total points. In sports an "error"—that is, not making the percentage play—is usually costly and only rarely is it luckily beneficial (a soccer shot that would miss the goal might luckily be deflected by a defender into the goal, for instance). But in the probability prediction context, predicting a 60% or 40% probability when the true probability is 50% is almost equally beneficial or costly. Loosely speaking, errors in predicting probabilities have only a second-order effect: 100 errors in a sequence of sports matches are more costly than 100 errors in predicting probabilities, and the latter might indeed by pure luck be overall beneficial.

3.2. Practical Relevance?

Our model is over-simplified in many ways; does it have implications for real-world prediction tournaments? Currently (announced February 2018) IARPA is offering $200,000 in prizes for top performers in its upcoming Geopolitical Forecasting Challenge (IARPA 2018); no doubt this will encourage volunteers to participate, but is it effective in identifying the best forecasters?

The authors of Tetlock, Mellers, and Scoblic (2017) write "some forecasters are, surprisingly consistently, better than others," and background to this assertion can be found in Mellers et al. (2015):

> [the winning strategy for teams over several successive tournaments was] culling off top performers each year and assigning them into elite teams of superforecasters. Defying expectations of regression toward the mean 2 years in a row, superforecasters maintained high accuracy across hundreds of questions and a wide array of topics.

Designers of that strategy were implicitly assuming that doing well in a tournament is strong evidence for ability (rather than luck), though our model results suggest that this assumption deserves some scrutiny. However, a main focus of the recent literature is arguing for the effectiveness of training methods, so that (if it were correct to downplay the effectiveness of "culling off top performers each year" in selecting for prior ability) our results actually reinforce that argument.

A superficial conclusion of our results is that winning a prediction tournament is strong evidence of superior ability *only* when the better forecasters' predictions are *not* reliably close to the true probabilities.[2] But are our models realistic enough to be meaningful? Two features of our "simple model for predictions by contestant with RMS error σ" are unrealistic. One is that contestants have no systematic bias toward too-high or too-low forecasts. A more serious issue is that the errors are assumed independent over both questions and contestants. In reality, if all contestants are making judgments on the same evidence, then (to the extent that relevant evidence is incompletely known) there is surely a tendency for most contestants to be biased in the same direction on any given question. Implicit in our model (and in our "first explanation" previously) is that, in a large tournament, this "independence of errors" assumption means that different contestants will explore somewhat uniformly over the space of possible prediction sequences close to the true probabilities, whereas in reality one imagines the deviations would be highly nonuniform.

[2]Asking whether "close to the true probabilities" is true in practice leads to basic issues in the philosophy of the meaning of *true probabilities*, not addressed here.

For recent relevant technical literature see Witkowski et al. (2017, 2018) and citations therein. In particular, when viewed as game theory with each player's only objective being to win the tournament, under the usual scoring scheme the optimal strategy involves *not* making truthful predictions, so one can study alternative scoring schemes that incentivize truthful reporting and are more likely to identify the bests forecasters.

Acknowledgments

I thank Seth Goldstein, Don Moore, Jens Witkowski, and two anonymous referees for valuable comments on an earlier draft.

Funding

The author gratefully acknowledges support by N.S.F. grant DMS-1504802.

References

Hung, K., and Fithian, W. (2019), "Rank Verification for Exponential Families," *The Annals of Statistics*, 47, 758–782. [5]

Hunter, D. R. (2004), "MM Algorithms for Generalized Bradley–Terry Models," *The Annals of Statistics*, 32, 384–406. [2]

IARPA (2018), "Geopolitical Forecasting Challenge Announcement," available at *https://www.herox.com/IARPAGFChallenge*. [5]

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., and Tetlock, P. (2015), "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions," *Perspectives on Psychological Science*, 10, 267–281. [5]

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., and Tetlock, P. E. (2014), "Psychological Strategies for Winning a Geopolitical Forecasting Tournament," *Psychological Science*, 25, 1106–1115. [1]

Tetlock, P. E., and Gardner, D. (2015), *Superforecasting: The Art and Science of Prediction*, New York: Crown. [1]

Tetlock, P. E., Mellers, B. A., and Scoblic, J. P. (2017), "Bringing Probability Judgments Into Policy Debates via Forecasting Tournaments," *Science*, 355, 481–483. [1,5]

Witkowski, J., Atanasov, P., Ungar, L. H., and Krause, A. (2017), "Proper Proxy Scoring Rules," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'17)*. [6]

Witkowski, J., Freeman, R., Vaughan, J. W., Pennock, D. M., and Krause, A. (2018), "Incentive-Compatible Forecasting Competitions," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*. [6]