A True Measurement of Ability in Track and Field

Jack D'Amico

CS 109 Challenge — August 2026

1 The problem

Every ranking for track and field ranks athletes by their PR, but that isn't actually the best measurement, since it ranks athletes not on their ability but on the best of however many attempts that athlete had. This makes published rankings systematically wrong in a way that probability can correct.

For example, two track and field athletes have the same season's best. One of them had 4 races and the other had 12. Every rankings site would rank these athletes as equal since their PRs are identical. But they aren't. This is because the athlete with 12 races had 3 times as many chances to get a good race day.

If each race is $T_{ij} = \mu_i + \varepsilon_{ij}$, then what is being ranked isn't μ_i, it's $\min_j T_{ij}$. The minimum of n draws gets further from the mean as n grows, so the recorded PR tends to overstate ability by an amount that depends on the race count. If we can invert that relationship and recover ability from the PR, provided we know n, then we have a better measurement of ability.

2 A probabilistic model of a personal record

In order to analyze the data we treat each race as an athlete's true ability plus a random race-day disturbance.

$$T_{ij} = \mu_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_e^2) \tag{1}$$

T_{ij} is athlete i's mark in race j. μ_i is their unobserved true ability. ε_{ij} is race-day conditions. σ_e is how much the day matters in event e.

Races are independent and identically distributed, which is important for the next step.

$$Y_i = \min_{j=1,\ldots,n_i} T_{ij} \tag{2}$$

This is what the rankings publish. Y_i is the minimum of all the athlete's races, which is the athlete's PR. But ability is μ_i, which isn't the same thing. It is a minimum because the shortest time is the best, which is why we negate field marks so that lower is better in every event.

With $z = (t - \mu_i)/\sigma_e$, and F and f the standard normal CDF and density:

$$P(Y_i > t) = \prod_{j=1}^{n} P(T_{ij} > t) = \big[1 - F(z)\big]^n \tag{3}$$

The PR can only be slower than t if every single race was slower than t, because the PR is the fastest race of the season. That turns one statement about the minimum into a statement about all n races at once. Because the races are independent we can multiply their probabilities together, and because they are identically distributed each of those probabilities is the same number, so the product becomes that number raised to the power n. The more races an athlete runs, the harder it is for all of them to be slow, which is what pulls the PR faster even when ability has not changed.

$$f_Y(t) = \frac{n}{\sigma_e}\big[1 - F(z)\big]^{n-1} f(z) \tag{4}$$

Differentiating the CDF gives the density of the PR, and the three factors can be read directly. For the PR to land exactly at t, one of the races has to land at t, which is the $f(z)$ term, and there are n choices of which race that was. Every other race has to come in slower, which is the $[1-F(z)]^{n-1}$ term. Setting $n = 1$ collapses this to the ordinary normal density.

With $z = (y_i - \mu)/\sigma_e$, the log-likelihood and its derivative are

$$\ell(\mu) = \log n - \log \sigma_e + (n-1)\log\big[1 - F(z)\big] + \log f(z) \tag{5}$$

$$\frac{d\ell}{d\mu} = \frac{1}{\sigma_e}\left[z + (n-1)\,\frac{f(z)}{1 - F(z)} \right] \tag{6}$$

Setting the derivative to zero and rearranging gives $z = -(n-1)\,f(z)/[1 - F(z)]$. The fraction is the hazard function, which is always positive. This means z will be negative. Since $z = (y_i - \mu)/\sigma_e$, that means $y_i < \hat{\mu}$, so the observed PR is faster than the estimated ability. This is the important distinction and what I am using to measure true ability. The estimator asks what ability, given n attempts, would typically produce a best mark of y_i. Since the best of several tries beats a single typical try, the ability is worse than the PR, and the $(n-1)$ factor pushes it further the more races were run. This results in estimated ability always being worse than the PR, and the gap growing as n grows. The correction's direction comes from this math output and is what is used to calculate true ability in our model.

There is no closed-form solution because the hazard term cannot be inverted, but the expression is monotone in μ, so the root is found numerically.

$$\mu_i \sim \mathcal{N}(\mu_{0e}, \tau_e^2), \qquad \log p(\mu \mid y_i) \;\propto\; \ell(\mu) - \frac{(\mu - \mu_{0e})^2}{2\tau_e^2} \tag{7}$$

There is a prior because 33% of the athlete-events have $n = 1$, where the likelihood is nearly flat.

For $n > 1$ the likelihood isn't quadratic in μ, so this is not the conjugate normal–normal update. MAP is found numerically.

Exact only at $n = 1$:

$$\hat{\mu}_{\text{MAP}} = \frac{\tau_e^{-2}\mu_{0e} + \sigma_e^{-2} y_i}{\tau_e^{-2} + \sigma_e^{-2}} \tag{8}$$

A precision-weighted average: whichever source is more precise pulls harder, so with one race the prior dominates. This is shrinkage.

3 Data and estimation

The data is scraped from TFRRS, which is an aggregation of all college track meet results. We collected the full men's rosters of all 17 ACC teams for the 2026 outdoor season, including both running and field events, with field marks negated so that lower is better. This gave 607 athletes and 20,491 marks, which reduce to 1,069 athlete-events across 511 athletes.

The race-day variance σ_e is estimated per event by pooling the within-athlete variance of every athlete with at least three marks in that event.

The prior parameters μ_{0e} and τ_e were fit from the sample means of athletes with $n \geq 4$ rather than from PRs. This matters because a PR carries the exact bias the model exists to correct, so fitting the prior on PRs would build that bias into the population mean and cancel out part of the correction.

Across the dataset, 33% of athlete-events have $n = 1$ and 55% have $n \leq 2$. For that many athletes the likelihood alone carries very little information, so the estimate has to lean on the prior.

4 Results

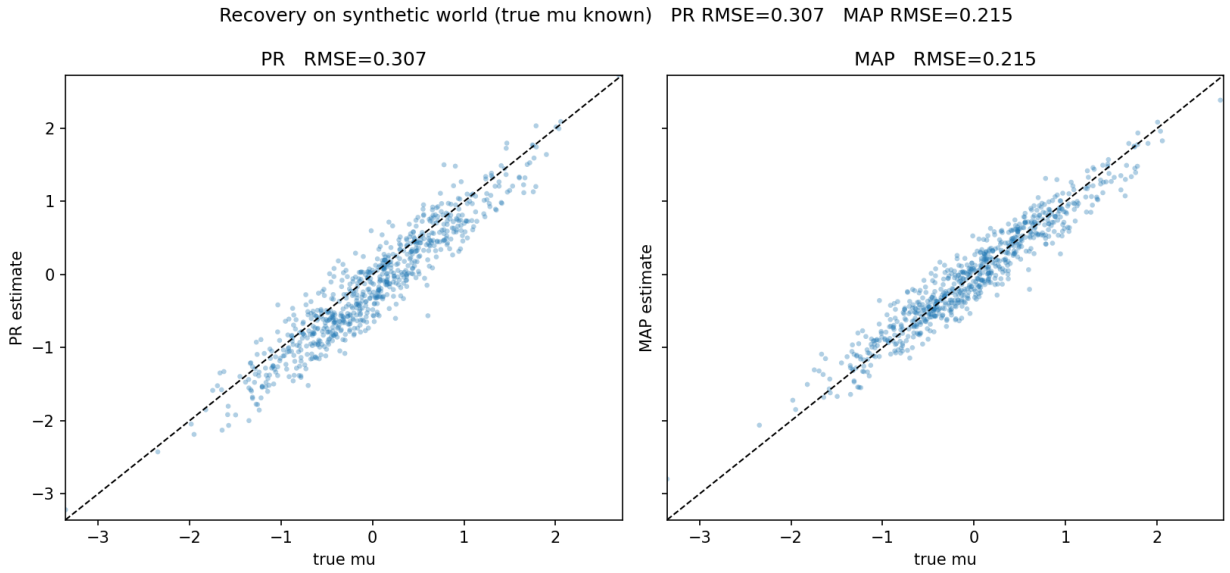

Figure 1: Recovery on synthetic data, where true μ is known. The PR (left) sits below the diagonal, crediting athletes with more ability than they have. MAP (right) recovers the diagonal. RMSE $0.307 \rightarrow 0.215$.

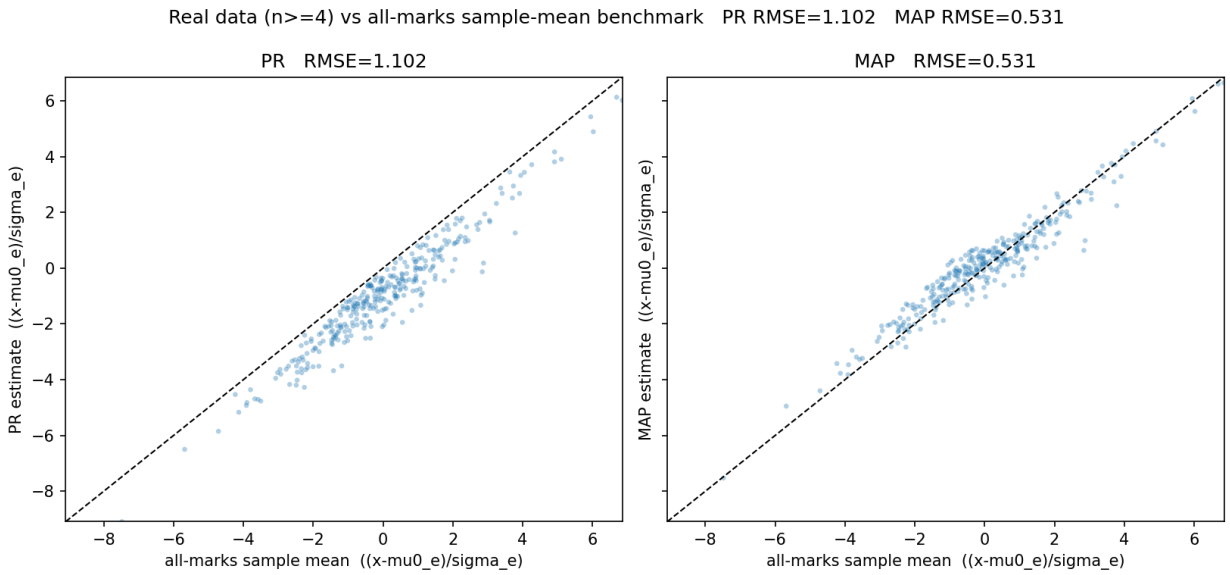

Figure 2: Real ACC data, compared against the sample mean of every mark an athlete recorded. Restricted to $n \geq 4$ so the benchmark is stable. The PR bias is visible without reading the numbers. RMSE 1.102 $\rightarrow$ 0.531.

On synthetic data MAP recovered ability better than the PR does (0.215 vs 0.307). On real ACC data the truth is unknown, but every race was scraped, so the all-marks sample mean serves as a stand-in, and the correction cuts error by more than half (0.531 vs 1.102).

The synthetic data shows the derivation and code are correct, but it was generated from the same model, so it wouldn't be an indication of the model fitting reality. However the real data shows that the model applies to real athletes and is valuable for measuring athlete ability.

5 Why this matters

Race count isn't random. Athletes at bigger programs have more meets and accumulate more attempts than equally talented athletes at smaller programs. So ranking on PRs doesn't just add noise, it systematically favors athletes who got more opportunities. The correction is a fairness fix as well as an accuracy fix.

6 Limitations

- Every athlete in the championship sample qualified for the meet, which is a selection on outcome and biases the estimates for low-n athletes upward.
- Strategic prelim rounds are run to qualify rather than to peak, so n counts races entered rather than maximum-effort attempts.
- Field marks are already a maximum over roughly six attempts within a meet, so the effective attempt count exceeds n and the correction under-adjusts field athletes relative to runners.
- The benchmark has its own noise. The sample mean's standard error at $n \geq 4$ is about $0.5\sigma_e$,

close to MAP's measured RMSE, so 0.531 is likely an upper bound on the true error.

- Ability is assumed constant across the season, ignoring fitness gains.
- Three sparse events had no athletes meeting the $n \geq 4$ threshold and fall back to global prior parameters.

7 Future work

This sort of model could be used to inform recruiting decisions for schools. It can measure each athlete on actual ability instead of PR, which can identify undervalued athletes. In the future I would try to simulate the conference championship on corrected abilities, compute each athlete's marginal expected points, and then value transfers under a fixed budget to try to build a winning team.

Use of LLMs

I used LLMs in order to create web scraping scripts for collecting athlete data. I also used them for planning the project and for idea generation. Lastly I used them for structuring the writeup with LaTeX typesetting and for the meet simulator. I worked through the derivation with LLM assistance. What I contributed was the idea, the execution of the project, and the writeup.