High-Stakes Performance Thresholds, Venue Choice, and the Market for Downhill Marathons*

Adam Looney†

Preliminary draft. Comments welcome.
March, 2026

Abstract

Performance cutoffs are widely used to allocate scarce opportunities, but when candidates are evaluated in environments that differ in difficulty, a common threshold confounds underlying ability with evaluation conditions, and creates incentives to select more favorable venues. In such settings, whether a candidate clears the threshold depends on the distributions of individual performance and environment effects, as well as on behavioral responses to the threshold. I study this problem in marathons, where runners compete for entry to the Boston Marathon using absolute qualifying times but can choose among courses that differ substantially in speed. Using 12.7 million finish times and panel data on 7.9 million runners, I estimate an AKM-style decomposition of performance into persistent runner effects and course effects. I then use these estimates to measure how tighter qualifying standards, growth in the supply of fast courses, and selective course choice by marginal runners have changed the composition of Boston qualifiers. The rising share of qualifiers from fast courses reflects all three forces: tighter standards increase the importance of venue differences in the thin tail of the performance distribution, the supply of fast courses has expanded, and runners near the qualifying margin sort toward faster courses. The results imply that penalties targeted at a small number of extreme courses are unlikely to address the central distortion. Venue-adjusted qualifying standards are a more coherent remedy. More broadly, the analysis applies to admissions, hiring, promotion, and other allocation systems that use common performance thresholds across heterogeneous evaluation environments.

Keywords: screening, performance thresholds, venue heterogeneity, AKM decomposition, marathon, sports economics
JEL Classification: D82, J24, L83, Z20

*I am grateful to colleagues at the University of Utah for helpful comments and suggestions. The author has no financial relationships or other potential conflicts of interest to disclose. This research received no external funding or research support.

†Marriner S. Eccles Institute, David Eccles School of Business, University of Utah. Email: adam.looney@eccles.utah.edu.

1 Introduction

Common performance thresholds are often used to allocate scarce opportunities, including admission, certification, promotion, and compensation. When performance is measured in environments that differ systematically in difficulty, however, a common cutoff does not rank underlying performance alone. It also rewards favorable evaluation conditions. As thresholds tighten, these differences matter more: a given environmental advantage may become increasingly decisive near the margin, and agents may respond by sorting into easier venues.

The Boston Marathon provides a useful setting for studying this problem. Entry is awarded on the basis of age- and gender-specific qualifying times that can be achieved at any certified marathon, even though marathon courses differ substantially in speed. Some purpose-built downhill races are much faster than Boston itself, while others are materially slower. Because the Boston Athletic Association has historically applied a common qualifying standard across these heterogeneous courses, the system offers a transparent example of how a fixed threshold can confound performance with evaluation conditions. As participation has grown and standards have tightened, the share of qualifiers coming from fast courses has risen sharply.

This paper studies that problem using the empirical framework of models that decompose outcomes into individual effects and environment-specific effects, as in matched worker-firm and teacher value-added models. Using 12.7 million finish times and panel data on 7.9 million runners, I estimate an AKM decomposition of marathon performance into persistent runner effects and course effects (Abowd, Kramarz and Margolis, 1999; Card, Heining and Kline, 2013). Rather than using that framework to net out sorting in order to recover venue quality, I use it to study how sorting across heterogeneous venues interacts with a common qualifying threshold. I then use these estimates to quantify the roles of tighter qualifying standards, growth in the supply of fast courses, and selective course choice in the rise of fast-course qualification. Strategic sorting exists, but much of the shift reflects the interaction

of tighter qualifying standards with large cross-course differences and the expansion of fast-course supply. In the thin tail of the performance distribution where the Boston cutoff binds, small course advantages have large effects on qualification. As a result, the qualifying rule increasingly selects runners on the basis of course choice as well as underlying ability.

Several features make this setting unusually informative. First, I observe the near-universe of marathon results from the United States and 18 other countries since 2004, with runners linked across races in panel data. Second, Boston qualification is determined by a sharp, publicly known cutoff, allowing me to study selection and behavior near the margin. Third, runners who qualify on different courses are later observed running Boston itself, providing a common-performance benchmark and a direct test of whether fast-course qualifying times reflect underlying ability or course-specific advantage. The setting is also policy-relevant: the Boston Athletic Association has publicly recognized the issue and will introduce downhill-based qualifying adjustments beginning in 2027.

This paper makes three contributions. First, I develop a comprehensive measure of course speed using an AKM decomposition in the matched employer-employee tradition (Abowd, Kramarz and Margolis, 1999; Card, Heining and Kline, 2013). I use the framework in reverse: rather than treating sorting as a nuisance in order to identify venue quality, I use the estimated runner and course effects to measure the extent and consequences of sorting itself. Course effects span more than 100 minutes across road marathons. Observable characteristics, including elevation, terrain, and climate, explain 85% of this variation, but net elevation drop—the margin targeted by the BAA's new rule—explains only 5%. In the spirit of Chetty, Friedman and Rockoff (2014*a*), I assess the external validity of the index using a cross-race portability test and estimate a portability coefficient of 0.93.

Second, I quantify the forces behind the growth in qualification from fast courses. The share of Boston qualifiers coming from courses more than 5 minutes faster than Boston rose from 19% in 2004 to 40% by 2019; among the most extreme courses—those more than 10 minutes faster—the share more than tripled from 9% to 28%. I show that most of this shift

reflects reallocation across course types—more runners competing at fast courses—rather than changes in the qualifying advantage of fast courses conditional on racing there. A fixed-population counterfactual implies that threshold tightening alone does little; a Kitagawa decomposition attributes nearly all of the increase to composition rather than within-type qualification rates; and a sequential reweighting that additionally controls for runner composition within course types confirms the pattern. Within-type runner composition contributes little to the shift, and the average quality of the fast-course field has not risen as those races expanded. Near-threshold runners are disproportionately represented on fast courses in every year, and this excess grows over time, consistent with the broadening of the fast-course market providing more sorting opportunities for marginal qualifiers.

Third, I present evidence of a behavioral response to the qualifying rule, though the design does not identify intent at the individual level. Among runners near the qualifying threshold, the covariance between runner and course effects becomes negative, and bunching at the cutoff is 24% greater on fast courses—patterns consistent with strategic course choice. A decomposition based on individual-level predicted finish times—a counterfactual object unavailable in the standard bunching framework of Saez (2010)—shows that the response operates primarily through course choice rather than differential effort. Runners do respond to the incentive, but much of the distortion is already embedded in the interaction between a common threshold and heterogeneous venues.

These results have direct policy implications. The BAA's 2027 downhill penalties target net elevation drop, even though that variable explains only 5% of cross-course speed variation. I propose an alternative: course-specific qualifying adjustments based on the AKM speed index. Under this approach, a course estimated to be 10 minutes faster than Boston would face a qualifying standard that is 10 minutes tighter. More broadly, the results show that proxy-based penalties aimed at a small number of extreme courses cannot resolve the core distortion. Venue-adjusted qualifying standards are the more coherent remedy.

Three identification limits should be noted at the outset. First, I do not observe intent: I

can show that runners near the qualifying margin disproportionately appear on fast courses, and that this pattern grows over time, but I cannot distinguish runners who deliberately sought a course advantage from those who chose a fast course for unrelated reasons and benefited incidentally. The behavioral evidence is therefore consistent with strategic sorting but does not establish it at the individual level. Second, the AKM decomposition imposes additive separability, so estimated course effects capture average speed differences rather than runner-specific match effects. Third, the policy counterfactuals are partial-equilibrium exercises that hold the course market fixed; in practice, venue-adjusted standards would alter demand across races. Section 6 discusses these limitations in greater detail.

This paper contributes to several literatures. The measurement strategy builds on the AKM matched employer-employee framework (Abowd, Kramarz and Margolis, 1999; Card, Heining and Kline, 2013), with Empirical Bayes shrinkage used to address limited-mobility bias (Andrews et al., 2008) and portability validation following Chetty, Friedman and Rockoff (2014*a*,*b*).

The venue-choice margin relates to the multitasking framework of Holmström and Milgrom (1991): when a principal screens on a single output dimension but agents can adjust behavior along multiple margins, the performance measure becomes distorted. The bunching results extend the notch literature (Saez, 2010; Kleven et al., 2011) to a setting in which the behavioral response operates through venue choice rather than the intensive margin, and in which runner fixed effects provide individual-level counterfactuals unavailable in the standard framework. The round-number bunching evidence complements Allen et al. (2017), who document reference-dependent pacing at salient time marks among marathon runners; the present paper distinguishes that generic pacing response from the BQ-specific venue-choice channel.

The strategic response to thresholds also parallels findings in education (Jacob and Levitt, 2003; Neal and Schanzenbach, 2010), the insight that performance metrics can distort the processes they are meant to measure (Campbell, 1979; Goodhart, 1984), and more gener-

ally the observation that agents facing bright-line rules may alter behavior at the margin (Lazear, 2006). More generally, the analysis applies wherever heterogeneous evaluation environments are mapped to a common performance threshold, including standardized testing across schools, promotions across sales territories, and professional certification across testing centers.

The paper proceeds as follows. Section 2 describes the institutional setting and develops a simple framework that generates three testable propositions. Section 3 presents the AKM estimation, validates the course speed index, and examines what drives course speed variation. Section 4 documents the reallocation of qualifiers as thresholds tighten and the fast-course market expands—the paper's core empirical contribution. Section 5 examines whether runners near the qualifying margin respond strategically through course choice. Section 6 evaluates policy counterfactuals. Section 7 concludes.

2 Institutional Setting and Framework

2.1 The Boston Marathon Qualifying System

The Boston Marathon is the world's oldest annual marathon and one of its most prestigious. Unlike most large-city marathons, Boston relies on time-based qualification rather than a general lottery: runners must achieve an age- and gender-specific qualifying time at a certified marathon within the designated window. That structure is central to the event's identity. As B.A.A. President and CEO Jack Fleming recently put it, runners "earn a spot on the starting line," underscoring that Boston is designed to reward demonstrated performance. The threshold is therefore high-stakes, clearly stated, and publicly known.

It has also become substantially more selective over time. For men aged 18–34, the posted qualifying standard tightened from 3:10 before 2013, to 3:05 for 2013–2019, to 3:00 for 2020–2025, and now stands at 2:55 for the 2026 and 2027 races. In practice, however, the effective threshold is even more selective because Boston admits a fixed field and, when

qualified applicants exceed available slots, accepts runners in order of their margin below the posted standard. As a result, the operative cutoff is the posted BQ time minus an additional "cutoff time," so the true threshold has fallen by more than the posted standard alone would suggest.

Three institutional features are central to the analysis. First, the qualifying system applies the same time cutoff across all certified marathons, regardless of course characteristics. A 3:00 finish at the steeply downhill REVEL Big Bear Marathon and a 3:00 finish at the hilly New York City Marathon are treated identically, even though they reflect different underlying performances. Second, qualification is backward-looking: runners qualify at race r in year t and then enter Boston in year $t+1$ or $t+2$. This creates a natural paired structure between the qualifying race and subsequent Boston performance that makes it possible to test how well qualifying times port across venues. Third, demand for Boston entry has persistently exceeded supply, leading to periodic tightening of posted standards and, since 2012, a buffer system that admits qualified runners in order of their margin below the posted standard. These changes increase the stringency of the effective cutoff and provide the key source of variation exploited in the analysis.

In 2025, the B.A.A. announced a new policy for the 2027 race: qualifying penalties of 5 minutes for courses with 1,500 to 2,999 feet of net elevation drop and 10 minutes for courses with 3,000 feet or more (Boston Athletic Association, 2025). This policy is the first departure from the historical common-cutoff rule. The change was explicitly framed by the B.A.A. as part of an effort to refine registration as demand has reached record levels, and public discussion of the rule often casts steep net-downhill qualification as an issue of fairness or even "cheating" rather than comparable performance across venues.[1] My analysis provides a broader interpretation: the underlying problem is not limited to a small set of

[1]In announcing the rule, B.A.A. President and CEO Jack Fleming described the downhill adjustments as "the next step in our evolving Boston Marathon registration process" as demand for Boston entry has reached record levels, while discussion among runners frequently frames steep downhill qualifiers as a way to "cheat" the system. See `https://www.baa.org/races/boston-marathon/qualify/` and `https://www.reddit.com/r/AdvancedRunning/comments/1aku0vv/pros_and_cons_of_the_tunnel_marathons/`.

extreme downhill races, but arises whenever a common qualifying threshold is applied across courses of differing speed, thereby rewarding runners on fast courses and penalizing similarly able runners on slow ones.

2.2 Data

The primary data consist of a web-scraped database of marathon results compiled from race websites, marathon result aggregators such as MarathonGuide.com (2024), race timing providers, and related online sources. The data cover approximately 979 courses in the United States and 18 other countries over 2004–2025, supplemented by official results from the Boston Marathon, the Chicago Marathon, and the Marine Corps Marathon. In total, the dataset contains approximately 12.7 million finish-time observations.

To link runners across races, I use an iterative entity resolution procedure based on name, gender, and year of birth. The first step requires exact matches on all three fields and links 91.5% of observations; subsequent steps allow for common nicknames, name abbreviations, and small year-of-birth discrepancies. This procedure yields 7.9 million unique runner identities. A robustness check restricting attention to exact matches produces a Spearman correlation of 0.999 with the baseline course effects, indicating that the linkage procedure does not drive the results. Appendix A provides further detail on both the entity resolution and time variable selection procedures.

For the AKM estimation, I require runners to appear at least twice within their geographic region and to have a known year of birth, and I trim the slowest 0.1% of finishers within each race-year. The resulting estimation sample contains 5.5 million observations from 1.5 million regional runners across 966 courses and 7,499 race-year cells.

2.3 Conceptual Framework

Consider an agent with baseline performance $a \sim F(a)$ who chooses effort $e \geq 0$ and an evaluation environment $v \in \mathcal{V}$. Observed performance is

$$y = a + \kappa(v) + e + \varepsilon,$$

where $\kappa(v)$ captures the systematic advantage or disadvantage of environment v, and ε is noise. The agent incurs cost $C(e) + D(v)$, where $C(\cdot)$ is convex in effort and $D(v)$ captures the cost of selecting environment v. A principal observes y and applies a common threshold rule: the agent qualifies if and only if $y \geq \bar{y}$ (cf. Baker, 1992).[2]

This setup delivers three propositions.

Proposition 1: Common thresholds confound person and place. Even absent strategic behavior—with effort fixed and venue choice unrelated to κ—a common cutoff rewards favorable environments. The qualification probability for an agent with baseline performance a at venue v is $\Pr(a + \kappa(v) + \varepsilon \geq \bar{y})$, which is increasing in $\kappa(v)$. As a result, equally capable agents qualify at different rates solely because they are evaluated in different environments. Favorable venues will therefore account for a disproportionate share of qualifiers, even mechanically. I test this implication in Section 3.

Proposition 2: Tightening the threshold amplifies venue differences. As $\bar{y}$ rises, qualification rates fall at all venues, but they fall faster at unfavorable venues. The reason is simply the shape of the performance distribution. When the cutoff lies in the middle of the distribution, many agents at both fast and slow venues qualify, so a venue advantage affects only a modest share of cases. But when the cutoff moves into the thin right tail, very few agents remain near the margin. In that region, a fixed venue advantage shifts much more probability mass from

[2]In the marathon setting, lower finish times are better, so qualifying requires $T \leq T^{BQ}$. The framework uses the equivalent formulation in which higher y is better and $\kappa(v) > 0$ denotes a favorable venue. In the empirical sections, course effects are estimated from $\ln T$, so a negative course effect corresponds to a fast course in this framework.

below the cutoff to above it. As a result, the same difference in $\kappa(v)$ matters much more at a stringent threshold than at a lenient one, and qualifiers become increasingly concentrated at high-κ venues. I test this implication in Section 4.

Proposition 3: Strategic response is concentrated among marginal candidates. When venue choice is endogenous, agents near the threshold—those who are close to qualifying absent additional effort or a more favorable venue—have the strongest incentive to seek high-κ environments or increase effort. Agents far above the threshold qualify regardless, while agents far below it remain unlikely to qualify even under favorable conditions. Behavioral responses are therefore concentrated near the cutoff, and their aggregate contribution to the reallocation of qualifiers across venues is limited by the mass of such marginal agents. I test this implication in Section 5.

The framework decomposes the rise in qualification from fast courses into three parts: (1) differences across courses even when nobody behaves strategically, (2) the enhancement of the effect of those differences when standards become tighter, (3) and runners' own strategic responses in where they race and how hard they try. The first two are mechanical: they arise even if runners do not change their behavior. Strategic course choice and extra effort can worsen the distortion, but they are not needed to generate it.

The exogenous mobility assumption typically used to identify AKM models is violated in this setting, since runners near the qualifying threshold systematically select faster courses. But that violation is itself part of the object of interest. The AKM decomposition provides the estimated runner and course effects needed to construct individual-level counterfactuals and to distinguish the contributions of market structure, threshold tightening, and individual course choice described in Propositions 1–3.

3 Measuring Course and Runner Effects

3.1 AKM Estimator

I estimate the following specification:

$$\ln T_{irt} = \alpha_i + \delta_r + \gamma_t + f(\text{age}_{it}, \text{male}_i) + g(W_{rt}) + \varepsilon_{irt} \tag{1}$$

where T_{irt} is the finish time of runner i in race r in year t, α_i is a runner fixed effect capturing persistent baseline performance, δ_r is a race fixed effect capturing the persistent speed of course r (including its typical climate), and γ_t is a year fixed effect capturing common trends across all races. The function $f(\cdot)$ is a quartic polynomial in age, centered at 40, fully interacted with gender. The term $g(W_{rt})$ is a quadratic in demeaned race-day temperature, $g(W_{rt}) = \beta_1 \widetilde{W}_{rt} + \beta_2 \widetilde{W}_{rt}^2$, where $\widetilde{W}_{rt}$ is the deviation of race-day morning temperature from the marathon's historical mean. Because mean temperature is absorbed by the course fixed effect δ_r, the weather controls capture only transitory race-year temperature shocks. Humidity, wind, and precipitation add no incremental explanatory power and are omitted.

To reduce false links in the entity resolution procedure, I define runner identities separately within broad geographic regions (U.S., Canada, UK/Ireland, Australia/New Zealand, and Other), so the same physical runner competing in both the U.S. and UK is assigned separate fixed effects in each region. The model is estimated using `reghdfe` (Correia, 2017), with standard errors clustered by regional runner.

The key assumption is additive separability: after accounting for runner ability and observables, courses shift log finish times by a common amount. This is a plausible approximation for road marathons, where features such as elevation profile, terrain, and typical weather conditions tend to affect all runners proportionally in time rather than creating highly runner-specific match effects.

A separate concern is estimation noise. In AKM settings, fixed effects for courses with few linked runners or weak mobility connections are estimated imprecisely, and limited mobility can attenuate the raw fixed-effect estimates (Andrews et al., 2008). I address this problem using Empirical Bayes shrinkage, which treats the estimated course effects as noisy signals of underlying course speed and pulls imprecise estimates toward the grand mean. Intuitively, the procedure leaves well-measured courses largely unchanged while shrinking sparsely connected or noisy courses more aggressively. This improves out-of-sample reliability and reduces the risk that extreme estimated course effects are driven by sampling variation rather than true differences in course speed.

The overall R^2 of equation (1) is 0.85. Table 1 reports a progressive specification that demonstrates the contribution of each set of fixed effects. Demographics alone explain 11% of log-time variation. Adding runner fixed effects raises R^2 to 0.81, reflecting the dominance of persistent performance differences. Race fixed effects contribute an additional 2 percentage points, and year effects plus weather controls add 1 more—modest in levels but substantively important, as these capture the course speed variation that is the object of interest. The weather controls are highly significant: a 10°F warmer-than-average race day slows finish times by 1.8%, or about 3.8 minutes at a 3:30 pace.

Table 1: Progressive specification: sources of variation in log finish time

	(1) Demographics	(2) + Runner FE	(3) + Race FE	(4) + Year FE + Weather
R^2	0.107	0.812	0.834	0.845
Adj. R^2	0.107	0.745	0.774	0.789
RMSE	0.198	0.106	0.099	0.096
Observations	5,987,040	5,476,985	5,476,985	5,476,985

Notes: Dependent variable is log finish time. Demographics are a quartic in age interacted with gender. Runner, race, and year fixed effects are added progressively. Column (4) adds year fixed effects and a quadratic in demeaned race-day temperature. Standard errors clustered by runner.

3.2 From Fixed Effects to a Course Speed Index

The course speed index is constructed from the estimated race fixed effects $\hat{\delta}_r$, which capture persistent differences in course speed after netting out runner ability, age-gender profiles, and race-year shocks. Because the model is estimated separately within each geographic region, these fixed effects are identified only up to a region-specific constant. I normalize the U.S. estimates by setting $\hat{\delta}_{\text{Boston}} = 0$ and calibrate the remaining regions using approximately 24,000 trusted bridge runners who appear in both the U.S. and another region. For each non-U.S. region, I use the median difference in runner fixed effects among bridge runners to recover a calibration constant that places course effects on the Boston-centered U.S. scale.

I then apply Empirical Bayes shrinkage to the estimated course effects to reduce noise from sparsely connected courses. Intuitively, the procedure treats each estimated $\hat{\delta}_r$ as a noisy signal of an underlying permanent course effect and pulls imprecisely estimated values toward the overall mean, with stronger shrinkage for courses estimated with greater variance (Morris, 1983; Andrews et al., 2008). Letting $\widehat{V}_r$ denote the estimated sampling variance of $\hat{\delta}_r$ and $\hat{\tau}^2$ the cross-course variance of the underlying effects, the shrinkage factor is

$$w_r = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \widehat{V}_r},$$

so the shrunk estimate is

$$\hat{\delta}_r^{shr} = w_r \hat{\delta}_r.$$

Courses with little information are therefore pulled more strongly toward zero, while well-measured courses are left largely unchanged. Courses with fewer than 500 observations are excluded from the final index. The resulting sample contains 518 courses.

To express the index in minutes, I convert the shrunk log-time effect into a level difference relative to Boston:

$$\Delta_r = \left(\exp(\hat{\delta}_r^{shr}) - 1\right) \times T^{BQ},$$

where T^{BQ} is the relevant qualifying time in minutes. Positive values indicate courses slower than Boston, and negative values indicate courses faster than Boston. I define the BQ handicap as $-\Delta_r$, which is the number of minutes by which the qualifying standard would need to be tightened to equalize course difficulty.

3.3 Course Speed Estimates

Figure 1 shows that course speed varies dramatically across marathons. Panel A highlights 25 benchmark races, spanning purpose-built fast courses such as REVEL Big Bear and BQ.2 Spring Chance, major city marathons, and slower destination races. The figure makes clear that some courses offer very large qualifying advantages relative to Boston. But the asymmetry in public discussion is also striking: while attention often focuses on the small set of extremely fast courses, the distribution also contains many slow courses, including several that are substantially slower than Boston and, in some cases, slower relative to Boston than the fastest courses are fast. For instance, the Marine Corps Marathon is estimated to be about 9 minutes slower than Boston, and the New York City Marathon about 5 minutes slower.

Panel B shows the full distribution of course effects. Most courses lie to the right of Boston, indicating that Boston is faster than the typical marathon in the sample. This means that most runners pursuing a Boston qualifier do so from courses that place them at a meaningful disadvantage.

Within each geographic region, the connected set encompasses all races in a single component, confirming that the fixed effects are jointly identified. Cross-region calibration links these components via bridge runners, producing a unified index. Robustness checks using nine alternative specifications—varying the MSE percentile cutoff, sample restrictions, and minimum race size—yield Spearman correlations above 0.98 with the baseline, and key race estimates are stable within 1–3 minutes across specifications (Appendix B).

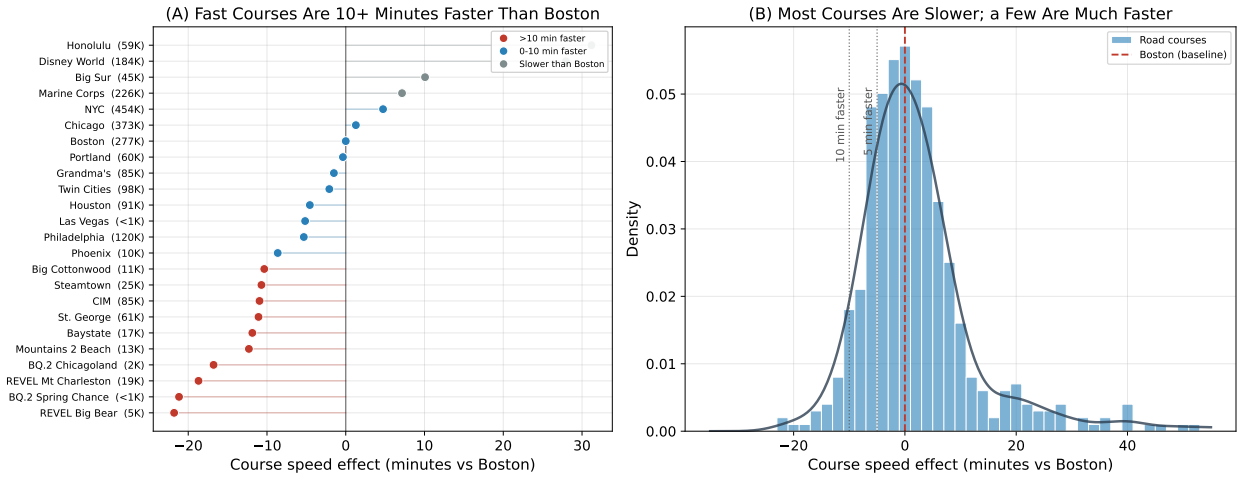


Figure 1: Course speed varies enormously across marathons. Panel A: course speed effects for 25 benchmark marathons (minutes relative to Boston). Panel B: full distribution of course speed effects for road marathons. Dashed line marks Boston (baseline).

3.4 What Drives Course Speed?

Course fixed effects are designed to capture everything that makes a marathon systematically faster or slower after accounting for runner ability and demographics. That is useful for measurement, but it also makes the estimates something of a black box: on their own, they do not reveal which underlying course features generate the observed differences. In this section, I open that black box by asking whether the estimated course effects line up with observable geographic and natural features, and in particular whether "downhill" is enough to explain the relevant variation. To do so, I collect detailed course profiles for the 320 courses with at least 500 observations and available profile data, covering elevation (gross gain, net drop, maximum elevation), course geometry (point-to-point, loop, out-and-back), surface (road, trail, mixed), climate (average race-day temperature), exposure (coastal), setting (urban, rural), and logistics (prize money, pacer programs). I then estimate cross-sectional regressions of the form:

$$\hat{\Delta}_r = \mathbf{X}_r'\boldsymbol{\beta} + u_r \tag{2}$$

where $\hat{\Delta}_r$ is the estimated course effect in minutes and $\mathbf{X}_r$ is a vector of observable course characteristics.

Table 2 reports three specifications of equation (2): net elevation drop alone, gross elevation gain alone, and a kitchen-sink model with all observables.

Table 2: Observable predictors of course speed: net drop misses most of the variation

	(1) Net drop only	(2) Gross gain only	(3) Kitchen sink
Net elev. drop (1000 ft)	-6.75^{***}		-4.99^{***}
	(0.85)		(1.14)
Gross elev. gain (1000 ft)		18.72^{***}	16.09^{***}
		(1.75)	(3.27)
Max elevation (1000 ft)			1.43^{*}
			(0.64)
Trail surface			10.82
			(6.52)
Avg. race temp (°F)			0.12
			(0.07)
N	320	518	320
R^2	0.047	0.768	0.852

Notes: Dependent variable is the course speed effect in minutes relative to Boston. Robust (HC1) standard errors in parentheses. Column (3) additionally includes course type, coastal, rural, prize money, and pacer program indicators. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Three findings stand out. First, net elevation drop—the margin targeted by the BAA's new rule—explains very little of the variation in course speed. Column (1) yields an R^2 of just 0.05. Although the point estimate implies that 1,000 feet of net drop makes a course about 6.8 minutes faster, most of the cross-course variation is unrelated to net elevation change. One reason is that many marathons begin and end near one another, so net drop is often modest even when the course itself is quite fast or quite difficult.

Second, gross elevation gain is far more important. Column (2) shows that cumulative climbing alone explains 77% of the variation in course speed, with each additional 1,000 feet of gain associated with roughly 18.7 minutes slower finish times. This is intuitive: a course with substantial climbing and descent can be much harder than a flat course even if both have the same net elevation change. Once gross gain is included, adding net drop contributes little additional explanatory power; in column (3), the R^2 rises only slightly, from 0.77 to 0.82.

Third, a broader set of observables explains most, but not all, of the remaining variation. Adding altitude, course type, surface, and climate raises the R^2 to 0.85. The estimated effects are also substantial: higher-altitude courses are slower, trail courses are dramatically slower, and warmer temperatures reduce performance. At the same time, 15% of the variation remains unexplained. That residual likely reflects factors such as crowding, wind exposure, crowd support, field quality, course layout, and other features that are difficult to summarize with simple observables but are still captured by the AKM course effect.

Appendix Figure 11 illustrates the contrast graphically using vigintile binscatters: net elevation drop explains almost nothing ($R^2 = 0.05$), while the observable difficulty index fits tightly ($R^2 = 0.85$). Appendix Figure 12 breaks down the estimated course effect for six representative marathons into contributions from each observable factor, illustrating why a formula-based correction cannot substitute for the data-driven approach.

3.4.1 General Cross-Race Portability

I construct a general paired dataset by linking each runner's destination race with their fastest prior marathon within 18 months, mirroring the window-based logic of BQ qualifying.[3] This yields 1.83 million pairs across 606,000 unique runners. I then estimate the following portability regression:

$$\ln T_{i,\text{dest}} - \ln T_{i,\text{orig}} = \beta \cdot (\hat{\delta}_{\text{dest}} - \hat{\delta}_{\text{orig}}) + \mathbf{X}_i'\boldsymbol{\phi} + \varepsilon_i \tag{3}$$

where $\hat{\delta}_{\text{dest}} - \hat{\delta}_{\text{orig}}$ is the difference in estimated course effects between the destination and origin races, and $\mathbf{X}_i$ includes changes in age controls and race-year controls. By differencing finish times within runner, this specification removes time-invariant runner ability and asks whether moving to a course estimated to be faster or slower predicts the corresponding change in performance. If the course index is perfectly portable, then $\beta = 1$: moving to

[3]The 18-month window matches the BAA qualifying window. Requiring different marathon names prevents self-matching. The "fastest prior" selection parallels how the BAA uses the fastest qualifying time.

a course that is 1 log point faster should improve finish time by 1 log point on average. Figure 2 presents the validation. Panel A shows the general portability binscatter for all 1.84 million pairs, while Panel B uses Boston as a standardized retest. Table 3 reports the corresponding regression estimates. The baseline coefficient on the FE difference is 0.93 (SE 0.003), indicating that course effects are highly portable across race pairs. Adding age controls or year fixed effects changes the estimate only modestly.

Table 3: Cross-race portability of course fixed effects

	(1) All pairs	(2) Boston dest.	(3) Non-Boston	(4) Wald test
FE difference	0.934***	0.835***	0.934***	
	(0.003)	(0.012)	(0.003)	
FE diff × Boston				0.837***
				(0.012)
FE diff × non-Boston				0.934***
				(0.003)
Age controls	✓	✓	✓	✓
N	1,834,265	111,399	1,722,866	1,834,265
R^2	0.140	0.097	0.147	0.147
Wald F-test				60.3

Notes: Dependent variable is the log time gap, $\ln T_{\text{dest}} - \ln T_{\text{origin}}$. "FE difference" is the course fixed effect at the destination minus the origin. Each runner's destination race is paired with their fastest prior race within 18 months (different marathon). Age controls include a quadratic in age and gender. Column (4) interacts the FE difference with a Boston-destination indicator; Wald F-test is for equality of the two FE-difference coefficients. Standard errors clustered by runner in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

3.4.2 Boston as a Standardized Retest

Boston provides a particularly useful validation case because it is a common destination: every runner in this exercise qualifies at some origin marathon and is then observed running the same race, Boston. Since Boston is the normalization point in the course index, $\hat{\delta}_{\text{Boston}} = 0$ by construction, it serves as a standardized retest of whether performance at the qualifying race carries over to a common venue. I therefore construct a paired dataset of runners who qualified at a non-Boston marathon within the BQ window and then ran Boston, yielding

111,399 qualifying-race/Boston pairs. On average, runners are slower at Boston than at their qualifying race, and 84.9% record a slower time at Boston.

The key question is whether that slowdown is systematically related to the speed of the qualifying course. If the course index captures real differences in course difficulty, then runners who qualified on faster courses should, on average, slow down more when they later run Boston. Figure 2 Panel B shows exactly this pattern: the Boston slowdown is larger for runners who qualified on courses with larger estimated speed advantages.

The general portability regression provides a direct way to quantify this relationship. As reported in equation (3), the dependent variable is the log time gap, $\ln T_{i,\text{dest}} - \ln T_{i,\text{orig}}$, and the key regressor is the corresponding gap in estimated course effects, $\hat{\delta}_{\text{dest}} - \hat{\delta}_{\text{orig}}$. A coefficient of $\beta = 1$ implies perfect portability: differences in estimated course effects fully predict differences in realized times across venues. Across all 1.83 million pairs, the baseline estimate is $\beta = 0.93$ (SE 0.003), indicating that course effects are highly portable. Restricting to Boston-destination pairs yields $\beta = 0.84$ (SE 0.012), compared with $\beta = 0.93$ (SE 0.003) for non-Boston destinations; a Wald test rejects equality ($p < 0.001$).

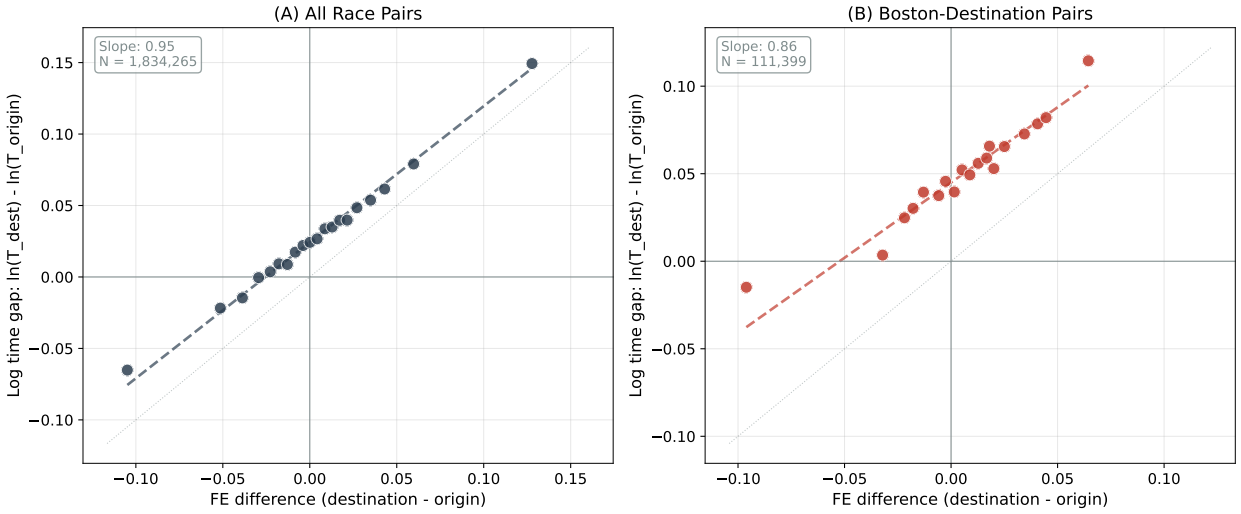


Figure 2: Course effects are portable across race pairs. Panel A: log time gap vs. FE difference for 1.83 million race pairs (slope = 0.93; $\beta = 1$ implies perfect portability). Panel B: Boston-destination pairs only (slope = 0.84).

The modest attenuation from unity likely reflects several forces. Measurement error

in the estimated course effects biases β toward zero. Runners qualify using their fastest recent marathon, which induces regression to the mean. Qualifying races may also involve unusually high effort or race-specific preparation that does not fully carry over. Crucially, every one of these forces pushes β below one, so 0.93 is a lower bound on portability, not a sign that the index overstates course speed. The non-Boston coefficient of 0.93 is especially informative because it is free of qualifying-pathway selection: these are general race pairs in which runners move between courses for any reason, and the course index predicts 93% of the resulting time change. The lower Boston-destination coefficient (0.84) is exactly what one would expect given the selection mechanisms documented in Section 5—fastest-prior-race selection, effort spikes near the qualifying threshold, and regression to the mean—and the 10-percentage-point gap between the two coefficients provides a rough upper bound on the combined magnitude of those forces.

A horse race between the course fixed effect and net elevation drop in predicting Boston slowdown confirms that the fixed effect is the stronger predictor: the marginal FE has a coefficient of -0.55 ($t = -4.9$), and when both are included, the FE remains highly significant while net elevation drop retains some independent predictive power (Appendix Table 6). Appendix J reports a complementary test: FE-adjusting qualifying times barely changes their predictive power for Boston performance ($R^2 = 0.545$ vs. 0.546), consistent with course effects capturing real speed differences rather than runner sorting.

3.5 Sorting in the AKM Decomposition

A useful feature of the AKM decomposition is that it separates runner ability, course difficulty, and the covariance between them. In standard applications, that covariance is a nuisance term. Here it is substantively important, because it captures whether stronger or weaker runners sort toward faster courses.

Table 4 reports the decomposition for the full sample and for subsamples defined by proximity to the BQ threshold. In the full sample, most variation in finish times is explained

by persistent runner differences, while course effects account for a smaller but still meaningful share. The key result, however, concerns the covariance term. In the full sample, the correlation between runner and course effects is modestly positive, suggesting that faster runners are somewhat more likely to run faster courses. But among runners near the BQ threshold, this pattern reverses: the correlation becomes negative, indicating that weaker runners sort into faster courses precisely where course choice is most valuable for qualification. Among runners far from the threshold, the correlation is close to zero. This pattern is consistent with strategic sorting being concentrated among marginal qualifiers.

Table 4: Runner-course sorting in the AKM decomposition

	Corr(runner, course)	Course share of variance	N
Full sample	+0.10	5.2%	4,583,946
Inframarginal ($>$ 30 min)	+0.01	10.4%	3,079,418
Near BQ ($\pm$10 min)	−0.09	11.4%	545,464
Qualifiers (0–5 min)	−0.10	11.6%	153,926

Notes: The table reports the correlation between runner and course fixed effects from the AKM decomposition, along with the share of variance in $\ln T$ explained by course effects. BQ margin is defined as qualifying standard minus finish time in minutes. The course share rises in the near-BQ subsamples because conditioning on a narrow finish-time band compresses total variance while course effects persist. The key result is that the runner-course correlation is modestly positive in the full sample, close to zero for inframarginal runners, and negative near the qualifying threshold, consistent with slower runners sorting into faster courses when qualification is at stake.

4 Threshold Tightening and the Reallocation of Qualifiers

The previous section showed that course speed differences are large, portable across race pairs, and closely related to observable terrain and climate features. Proposition 2 implies that as the qualifying threshold tightens, these course differences should matter more: even holding runner behavior fixed, stricter standards will shift qualification toward faster courses. This section documents that reallocation and examines how it reflects threshold tightening, changes in where runners compete, and behavior near the qualifying margin.

The analysis uses the individual-level counterfactual implied by the runner fixed effects. For each runner i at age a and gender g, I define the predicted finish time at course r as

$$\hat{T}_{ir} = \exp\big(\hat{\alpha}_i + \hat{\delta}_r^{shr} + \hat{f}(a, g)\big), \tag{4}$$

and the corresponding neutral-course baseline as

$$\hat{T}_i^0 = \exp\big(\hat{\alpha}_i + \hat{f}(a, g)\big),$$

which is the time runner i is predicted to run on a course with zero speed effect, normalized to Boston. These counterfactual predictions make it possible to distinguish the role of tighter qualifying standards from changes in where runners compete, and to study whether runners near the margin become increasingly concentrated on faster courses. This runner-level approach replaces the smooth counterfactual density used in standard bunching analyses (Saez, 2010) and underlies the decompositions in this section and the next.

4.1 The Shrinking Qualifying Menu

The amplification mechanism can be stated more concretely in runner-level terms. Each runner's neutral-course baseline performance—their predicted finish time on a course with zero speed effect—determines the set of courses on which they could qualify. As BQ standards tighten, that set shrinks sharply for runners near the margin.

Figure 3 Panel A illustrates this compression. A runner with a neutral-course baseline of 195 minutes could qualify at 57 courses under the pre-2013 standard of 3:10, but only 13 courses under the 2013–2019 standard of 3:05, and just 4 under the current 3:00 standard. More generally, the qualifying menu collapses rapidly as baseline performance worsens: moving from a 5-minute cushion to a 15-minute cushion reduces the number of qualifying options from over 400 courses to fewer than 10. This reflects the shape of the course distribution. Of the 518 indexed courses, 223 are at least slightly faster than Boston, but only 87 are more

than 5 minutes faster and just 33 are more than 10 minutes faster. As a result, tighter standards do not just reduce qualification rates overall; they disproportionately remove slower courses from the set on which marginal runners can still qualify.

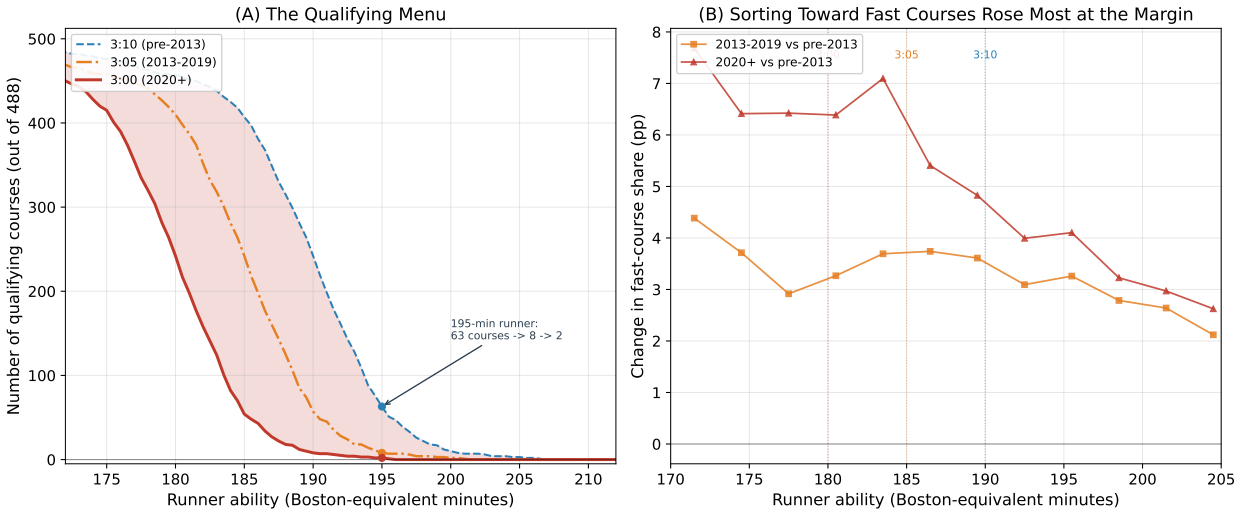


Figure 3: The shrinking qualifying menu. Panel A: number of qualifying courses by runner baseline performance under three BQ standards. Shaded region: squeeze zone. Panel B: change in the share of runners choosing a fast course (FE < 0) relative to the pre-2013 era, by baseline performance. The increase peaks near the BQ threshold, where the qualifying menu is most compressed.

Panel B shows the behavioral counterpart to this compression. I plot the change in the share of runners choosing a fast course (FE < 0) relative to the pre-2013 baseline, by neutral-course baseline performance. The increase in fast-course sorting is present across the ability distribution—reflecting the overall growth in the fast-course market—but it peaks sharply near the BQ threshold. In the 2020+ period, the increase reaches 10 percentage points at the 3:00 BQ margin (180 minutes), compared with roughly 7 percentage points for runners 20 minutes slower. The 2013–2019 period shows a similar pattern, peaking near the 3:05 threshold. This concentration at the margin is consistent with strategic sorting: runners whose qualifying menu has been most compressed are the ones most likely to shift toward faster courses.

The next subsections show that this compression helps explain the aggregate reallocation of Boston qualifiers toward fast courses, although most of that shift reflects changes in

participation across course types rather than changes in average runner composition within them.

4.2 The Rise of Fast-Course Qualification

Qualification from fast courses has risen substantially over time. Figure 4 Panel A shows that the share of BQ qualifiers coming from courses more than 5 minutes faster than Boston rose from 19% in 2004 to 41% by 2019, before settling at 34% in 2023. Among the most extreme courses—those more than 10 minutes faster—the share more than tripled from 9% to 28% over the same period. Participation at fast courses also increased: the share of all finishers at such courses rose from 10% in 2004 to 20% in 2019. But the increase among qualifiers was much larger, indicating that the growth in fast-course qualification reflects more than general market expansion alone. This period also coincides with the emergence of purpose-built qualifying brands (REVEL launched in 2013, and BQ.2 events explicitly market sub-three-hour qualifying potential) and with the BAA's periodic tightenings of qualifying standards in 2013 and 2020. Appendix Figure 13 shows that the fast-course share increases following both tightenings. The next subsections ask how much of this shift reflects threshold tightening interacting with course heterogeneity, how much reflects reallocation across course types, and whether runners near the qualifying margin become increasingly concentrated on faster courses.

4.3 Fixed-Population Counterfactual

To separate the role of tighter qualifying standards from changes in the marathon market, I construct three counterfactual scenarios.

First, in a fixed-population counterfactual, I hold fixed the observed distribution of runners, course choices, and finish times at their 2004–2009 values, and then apply each year's actual BQ standards to this fixed field. This asks how much the tightening of standards alone would have changed the composition of qualifiers if the marathon market had not evolved.

Second, in a fixed-threshold counterfactual, I allow the runner population, course choices, course entry, and finish times to evolve as observed, but I hold BQ standards fixed at their 2004 levels. This asks how much of the rise in fast-course qualification can be explained by changes in the marathon market absent any tightening of the qualifying rule.

Third, I construct a random-assignment counterfactual. For each year, I take the actual population of runners and strip their course fixed effects to recover each runner's latent ability. I then randomly assign these runners to courses using the 2004–2009 course-size distribution and apply the year's actual BQ standards. This counterfactual asks what would happen if runners chose courses without regard to qualifying considerations.

Figure 4 Panel B reports the results. Under the fixed-population counterfactual, the fast-course share among qualifiers remains roughly stable between 23% and 25%, indicating that tighter standards alone do little when the set of runners and courses is held fixed. Under the fixed-threshold counterfactual, the fast-course share rises from 25% to 39% by 2019, closely tracking the observed increase to 40%. The random-assignment counterfactual also stays flat, confirming that without strategic course selection, neither tighter thresholds nor changes in the runner population affect the fast-course share among qualifiers.

Comparing the fixed-population and random-assignment counterfactuals provides additional evidence of strategic sorting. Under random assignment, tighter standards should slightly *increase* the fast-course share, because runners at slow courses are deeper in the tail of the ability distribution and thus more likely to be eliminated as the threshold tightens. In the actual data, however, the fixed-population counterfactual shows a slight *decrease*, because fast courses have excess mass of runners bunched right at the qualifying threshold—exactly the runners who are swept out by tightening. This reversal is consistent with the strategic BQ-targeting behavior documented in Section 5.

The rise in fast-course qualification therefore does not appear to be driven by threshold tightening in isolation. Instead, it reflects changes in the marathon market—including the expansion of fast courses and changes in where runners compete—with tighter standards

magnifying the importance of those differences near the qualifying margin.

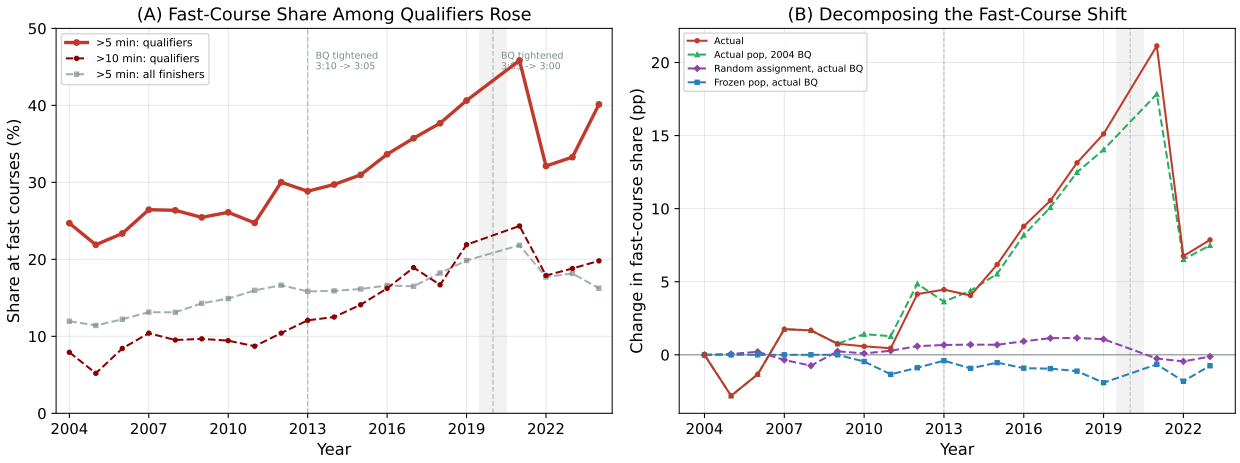


Figure 4: Tighter standards shifted qualification toward fast courses. Panel A: share of BQ qualifiers and all finishers at fast courses (more than 5 minutes faster than Boston), 2004–2024. Bold line: qualifier share at courses more than 5 minutes faster; thin line: qualifier share at courses more than 10 minutes faster; dashed line: all-finisher share at courses more than 5 minutes faster. Panel B: counterfactual decomposition. The actual fast-course share among qualifiers (solid) rises roughly 15 percentage points by 2019. The fixed-threshold counterfactual (actual population, 2004 BQ; dashed green) closely tracks the actual series, indicating that market expansion drives the shift. The random-assignment counterfactual (runners' abilities randomly assigned to 2004–2009 course distribution, actual BQ; dashed purple) and the fixed-population counterfactual (2004–2009 runners, actual BQ; dashed blue) both stay flat, confirming that neither threshold tightening nor changing runner composition moves the fast-course share absent strategic course selection.

4.4 Kitagawa Decomposition

Table 5 brings together three complementary decompositions of the rise in fast-course qualification. I begin with a Kitagawa decomposition that separates changes in the fast-course share among qualifiers into a composition component and a qualification-rate component, equation (5):

$$\Delta(\text{fast share}) = \underbrace{\Delta s_{\text{fast}} \cdot \bar{q}_{\text{ratio}}}_{\text{composition}} + \underbrace{\bar{s}_{\text{fast}} \cdot \Delta q_{\text{ratio}}}_{\text{rate}} + \text{interaction}, \tag{5}$$

where s_{fast} is the share of all finishers at fast courses and q_{ratio} is the relative qualification rate at fast versus other courses. The composition term captures changes in where runners race,

while the rate term captures changes in the qualifying advantage of fast courses conditional on racing there.

Panel A of Table 5 summarizes the raw time-series patterns and the counterfactuals. The actual fast-course share among qualifiers rose from 18.9% in 2004 to a peak of 39.9% in 2019, before falling to 33.5% in 2023. The fixed-population counterfactual remained between 18% and 20%, and the random-assignment counterfactual stayed near 20% throughout. This reinforces the conclusion that tightening standards alone, applied to a fixed set of runners and courses, does little to change the composition of qualifiers—and that without strategic course selection, neither threshold changes nor evolving runner populations move the fast-course share.

Panel B shows that most of the observed increase reflects composition rather than changes in relative qualification rates. By 2019, the total change in the fast-course share among qualifiers was 19.8 percentage points; of this, 19.0 percentage points reflect the composition effect—more runners competing at fast courses—while the rate effect contributes just 0.5 percentage points. In other words, the rise in fast-course qualification stems primarily from more runners appearing at fast courses, not from a growing qualification premium conditional on racing there. The underlying inputs point in the same direction: the share of all finishers at fast courses rose from 10% in 2004 to 20% in 2019, while the qualification-rate ratio between fast and other courses remained fairly stable, generally fluctuating between 2 and 3.

Panel C turns from qualifier shares to qualifier counts. Using an Oaxaca-Blinder-style sequential reweighting, I decompose the increase in the total number of qualifiers into a component associated with changes in course selection and a residual component associated with within-course changes in observed performance. Course selection contributes 18,536 additional qualifiers relative to 2004, while within-course performance shifts contribute 8,740. Although this decomposition addresses a different outcome, it points to the same conclusion as Panels A and B: changes in where runners compete account for a substantial share of

the growth in qualification. Taken together, the three decompositions imply that the rise in fast-course qualification is driven mainly by reallocation across courses, with tighter standards mattering through their interaction with that evolving course distribution rather than through a large change in conditional qualification rates.

A sequential reweighting that controls for changes in runner composition within course types reinforces this conclusion. Among runners with estimated fixed effects, the fast-course qualifier share rises by 20.4 percentage points relative to 2004 by 2019. Of this, 20.2 points are explained by reallocation across course types, −1.5 points by changes in within-type runner composition, and the remaining 1.7 points by a residual reflecting within-cell changes in qualification rates. The negative within-type term indicates that the average quality of the fast-course field has not risen and may have fallen slightly as those races expanded.

Table 5: Decomposition of the growth in fast-course qualification

	2004	2013	2019	2023
Panel A: Fast-course share among qualifiers (%)				
Actual	24.7	29.2	39.8	32.6
CF1: Frozen population, actual BQ	24.8	24.4	22.9	24.1
CF2: Actual population, 2004 BQ	24.7	28.3	38.7	32.2
CF3: Random assignment, actual BQ	20.0	20.7	21.1	19.9
Panel B: Kitagawa decomposition (pp change vs 2004)				
Composition (more runners at fast courses)	0.0	6.5	12.7	10.2
Rate (changing relative qualification rate)	0.0	-2.5	1.1	-2.2
Interaction	0.0	-0.4	0.3	-0.5
Total change	0.0	3.6	14.0	7.5
Panel C: Sequential reweighting (change in qualifiers vs 2004)				
Course selection	0	6765	14,908	18,536
Performance (within-course)	0	16,535	4016	8740
Demographics	0	560	530	140
Total change	0	23,861	19,454	27,417

Notes: Panel A reports the share of BQ qualifiers from fast courses (more than 5 minutes faster than Boston) under three scenarios: actual data, CF1 (frozen 2004–2009 population with actual BQ standards), and CF2 (actual population with fixed 2004 BQ standards). Panel B decomposes the change in fast-course qualifier share into composition (share of finishers at fast courses), rate (relative qualification rate), and interaction. Panel C reports an Oaxaca-Blinder sequential reweighting of the change in total qualifiers into course selection, within-course performance, and demographic channels. Components may not sum exactly due to rounding.

The detailed time-series versions of these decompositions are reported in Appendix E (Figures 14, 15, 16, and 17).

A related implication is that tighter standards also increase the share of qualifiers whose success depends on course choice. Among all qualifiers, the share who are course-dependent—that is, runners who qualify at their chosen course but whose Boston-equivalent time (stripping the course speed advantage) would not qualify—rises from 8% in 2004 to 17% by 2019.

Taken together, the decompositions in this section indicate that the growth in fast-course qualification is driven primarily by the interaction of tighter standards with an expanding fast-course market. These aggregate patterns do not by themselves establish strategic course choice, but they point to a natural next question: whether runners near the qualifying margin are disproportionately represented on fast courses, and whether that excess grows over time. Section 5 takes up that question directly.

5 Behavioral Response

The previous section showed that tighter standards shift qualification toward faster courses even if runners do not change their behavior. This section asks whether runners also respond strategically, and if so, whether that response operates mainly through course choice or through unusually strong performance conditional on course choice. Throughout this section, I interpret the patterns as consistent with strategic course choice, while recognizing that I cannot observe intent directly: runners near the margin who appear on fast courses may have selected those venues for reasons unrelated to qualification, such as geography or scheduling. Distinguishing these margins matters for interpretation. If bunching arises primarily because runners sort into faster venues, then the threshold mainly distorts where runners compete. If it arises primarily because runners outperform their usual baseline on a given course, then the threshold is associated with additional race-specific performance conditional on venue choice.

The empirical strategy combines bunching evidence with runner-level counterfactuals and direct tests of venue choice. Bunching at the BQ cutoff establishes that runners respond to the incentive, but bunching alone cannot reveal whether that response reflects selection into faster courses or unusually strong performance conditional on course choice. To separate those channels, I use the AKM decomposition to construct each runner's neutral-course baseline and predicted time at the chosen course. I then compare actual outcomes to these counterfactual benchmarks and supplement the bunching evidence with descriptive and regression-based tests of whether runners near the qualifying margin are disproportionately represented on faster courses. The analysis proceeds in three steps. First, I document bunching at the cutoff. Second, I show that runners near the margin sort toward faster courses. Third, I decompose excess qualification into a course-selection component and a residual performance component, which I interpret as effort in a broad sense.

5.1 Bunching at the Qualifying Cutoff

Pooling all age-gender groups by centering each observation on its own BQ standard reveals clear bunching at the qualifying cutoff, analogous to the density test of McCrary (2008). On fast courses, there is visible excess mass just above the threshold: the density in the 0–5 minute window above the cutoff is 24% higher than in the 0–5 minute window below, for a bunching ratio of 1.24. On other courses, the corresponding ratio is only 1.04. The same pattern appears at finer resolution. The 1-minute cliff ratio at the BQ mark is 1.24 on fast courses, compared with 1.13 on other courses (Appendix Figure 18). The qualifying threshold therefore stands out sharply from the surrounding distribution, especially on fast courses, indicating that runners do respond to the incentive.[4]

[4] A moving-BQ test exploits the M18–34 threshold shifts (3:10 → 3:05 → 3:00): the cliff ratio spike at 3:10 drops from 1.23 to 0.93 after 2013, confirming that the excess mass is specific to the BQ threshold rather than to round-number finish times. The underlying distribution of runner baseline performance is smooth across round-number marks, further supporting a behavioral interpretation. See Appendix H for additional bunching evidence, including the moving-BQ test (Figure 26), era-by-era decomposition (Figure 25), and the contrast between smooth baseline ability and spiky actual finish times (Figure 27).

5.2 Marginal Qualifiers and Course Sorting

Bunching alone does not reveal the margin of response. To assess whether it reflects strategic venue choice, I assign each non-Boston observation a BQ margin, defined as the qualifying standard minus finish time, so that positive values indicate qualification, and compare course choices across the margin distribution.

Figure 5 shows clear evidence of sorting into faster courses near the threshold. Panel A plots average course speed against the BQ margin. Mean course speed rises steadily as runners approach the cutoff, and there is a visible jump at the threshold itself: runners who just qualify are on courses averaging 4.7 minutes faster than Boston, compared with 3.9 minutes faster for those who just miss. By contrast, runners far from qualification (margin -30) are on courses only 2.4 minutes faster than Boston. Panel B shows the same pattern for the share of runners on very fast courses (FE < -10 minutes), which rises discretely by 3.9 percentage points at the cutoff.

The cross-sectional pattern is especially visible at purpose-built fast courses. BQ.2 Spring Chance sends 29% of its finishers to Boston, 6.4 times the baseline qualification rate, while REVEL Mt. Charleston sends 29%, 5.5 times the baseline rate. Large races such as Chicago and New York City send more runners in absolute terms, but at rates much closer to the population average. The evidence therefore suggests that runners near the margin are disproportionately represented on courses that offer an unusually large speed advantage.

The overrepresentation of near-threshold runners on fast courses is present throughout the sample period and grows over time. Using runner fixed effects to construct FE-adjusted Boston-equivalent finish times, I identify runners whose predicted performance lies within ± 5 minutes of the 2004 BQ standard. In 2004, 21.0% of these near-margin runners were on fast courses, compared with 14.2% overall; by 2017, the near-margin share had risen to 32.7% versus 18.3% overall; and in 2019 it reached 35.9% versus 23.5%. The growth in this excess is consistent with the broadening of the fast-course market providing more sorting opportunities for runners near the qualifying threshold, though it does not by itself identify

whether the sorting is caused by the threshold or by other factors correlated with proximity to it.

Within-runner panel evidence around the 2013 tightening provides a more direct test that this pattern reflects behavior rather than pure composition. Following the tightening from 3:10 to 3:05 for M18–34, runners in the squeeze zone (baseline 185–195 minutes) shifted toward faster courses by 0.4 percentage points more than runners far from the cutoff, consistent with substitution toward faster venues as the qualifying menu contracted.[5]

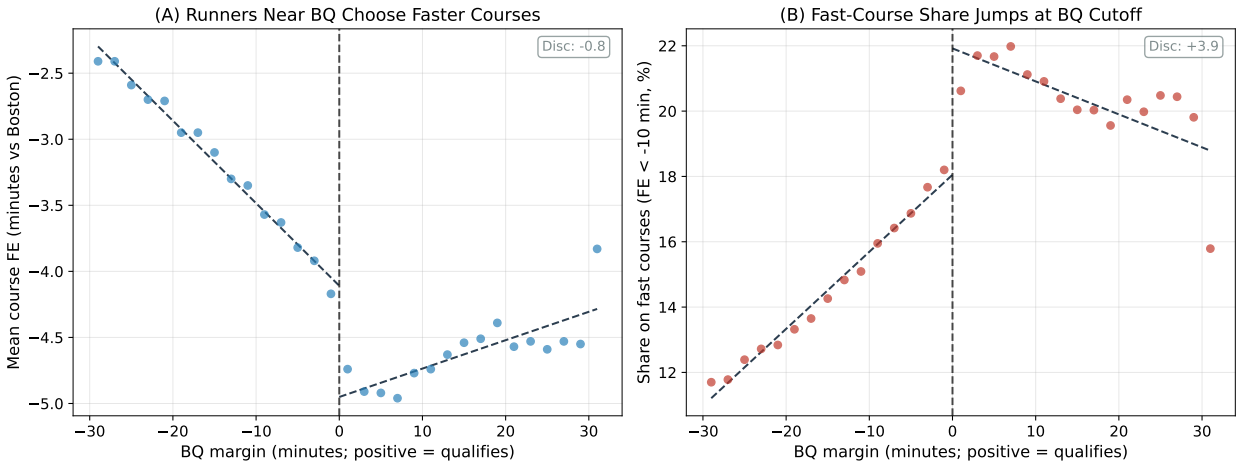


Figure 5: Marginal runners sort toward faster courses. Panel A: mean course FE (minutes) by 2-minute BQ margin bins. Panel B: share on fast courses (FE < -10 min). Positive margin = qualifies. Dashed lines show separate OLS fits above and below the cutoff.

5.3 Near-Miss Behavioral Response

Appendix Figure 19 provides additional evidence that the BQ threshold is behaviorally consequential. The figure plots three outcomes against the BQ margin. Near-misses and near-qualifiers retry at similar rates, around 50%, indicating that both groups remain engaged in the qualifying process. Course choice, however, differs: runners who just miss are somewhat more likely to choose fast courses than runners who just qualify, consistent with

[5]The panel comparison uses 36,425 M18–34 runners observed in both 2010–2012 and 2013–2015. The difference-in-differences estimate, defined as the change for squeeze-zone runners relative to runners slower than 200 minutes, is 0.20 percentage points in course FE, or about 0.5 minutes in predicted time.

continued search for venue advantage among those who remain below the threshold. The clearest discontinuity is in the probability of running Boston itself, which jumps by about 5.4 percentage points at the cutoff. This confirms that the threshold is binding and that clearing it meaningfully changes subsequent behavior.

5.4 Effort Decomposition

Bunching and course sorting show that runners respond to the qualifying incentive, but they do not by themselves reveal whether that response operates mainly through venue choice or through unusually strong performance conditional on course choice. I therefore use the individual-level counterfactual from equation (4) to separate these channels.

For each observation, I compute the runner's predicted finish time using the estimated runner effect, course effect, and age-gender profile. The predicted distribution is smooth through both round-number and BQ thresholds, confirming that the bunching in actual times is not mechanically built into the fixed effects (Appendix Figure 21). I then define the residual

$$r_{ir} = \ln T_{ir} - \ln \hat{T}_{ir},$$

which measures residual performance relative to the model's prediction at the chosen course. I interpret this residual as effort only in a broad sense: it may also capture race-day conditions, pacing, or other transitory factors not absorbed by the fixed effects. If bunching at the BQ cutoff reflects extra performance conditional on venue, that pattern should appear in this residual. If it reflects course selection instead, the residual response should be limited once the course effect is accounted for.

A useful benchmark comes from ordinary round-number goals such as 4:00 and 3:30, where runners may pace strategically but course speed does not create a Boston-specific qualification advantage. At those thresholds, the residual response looks similar on fast and other courses, suggesting that generic pacing or motivational spikes are not enough to

explain the distinctive bunching pattern at BQ cutoffs. By contrast, at the BQ threshold the residual patterns diverge by course type (Appendix Figure 22). The remaining question is whether that residual difference is large enough to account for the observed bunching, or whether most of the response still operates through venue choice.

Figure 6 summarizes this comparison in a simple 2×2 design: {BQ threshold, round-number threshold 4:00} $\times$ {fast courses, other courses}. At the 4:00 threshold, cliff ratios are similar across course types, 1.55 on fast courses versus 1.45 on other courses. At the BQ threshold, they diverge more sharply, 1.38 on fast courses versus 1.11 on other courses. The resulting difference-in-differences of 0.17 captures the excess bunching at the BQ cutoff on fast courses beyond what generic pacing behavior would predict. This pattern suggests that venue choice is an important part of the behavioral response, while any additional residual-performance effect appears smaller.

A moving-BQ calibration using the 2013 and 2020 threshold shifts provides a further check. The BQ stakes premium is 0.30 at 3:10 and 0.22 at 3:05, showing that the bunching moves with the qualifying standard itself rather than simply attaching to salient round numbers. Taken together, these results indicate that runners do respond behaviorally to the BQ incentive, but that the response appears to operate primarily through venue choice rather than through large increases in performance conditional on course.

Do fast-course qualifiers rely less on residual performance because fast courses already provide a speed advantage? Among marginal qualifiers (BQ margin 0–5 minutes), the mean residual is -0.044 log points on fast courses, compared with -0.058 on other courses, a difference of 0.015 log points, or about 1.5% of predicted time. Fast-course qualifiers therefore outperform their predicted baseline by slightly less than comparable qualifiers on other courses. This pattern is consistent with venue choice and residual performance acting as partial substitutes. Quantitatively, however, the residual gap is small relative to the typical 5–10% speed advantage of fast courses, indicating that the behavioral response operates primarily through venue choice rather than through large differences in performance conditional

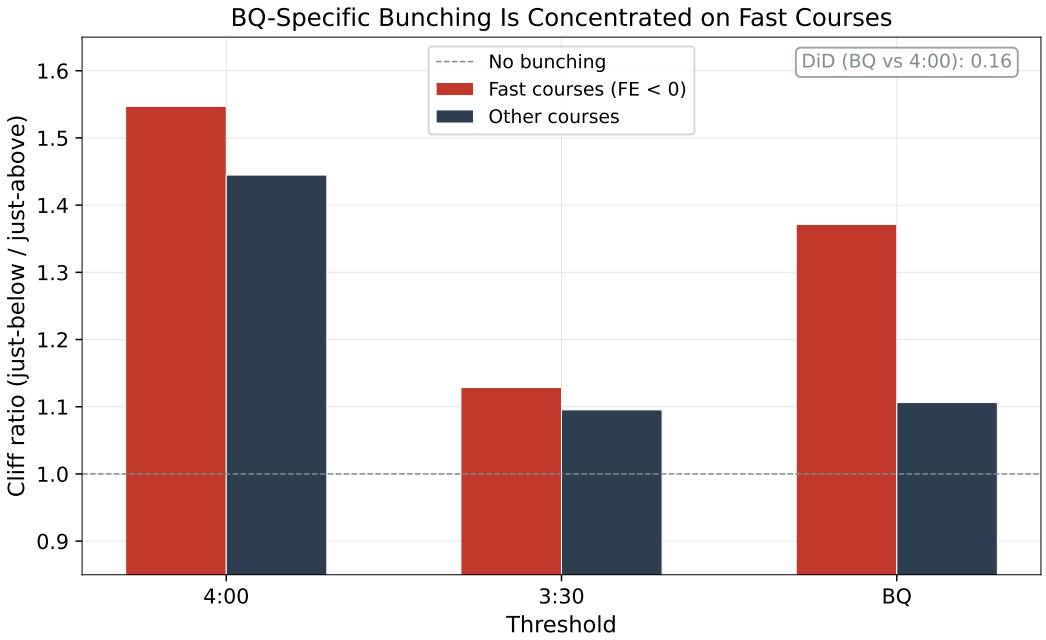


Figure 6: BQ-specific bunching is concentrated on fast courses. At 4:00 and 3:30, cliff ratios are similar across course types. At the BQ threshold, they diverge more sharply (fast = 1.38, other = 1.11). The difference-in-differences relative to 4:00 is 0.17.

on course.

Appendix I reports a complementary natural experiment using weather shocks: when unusually hot conditions cause near-BQ runners to miss the qualifying threshold, those runners do not differentially switch to fast courses in their next attempt, consistent with substantial inertia in course choice.

6 Policy Counterfactuals

6.1 The Case Against Proxy-Based Corrections

The B.A.A.'s 2027 policy assigns qualifying penalties based on net elevation drop alone. That approach has an intuitive appeal, but it is a poor proxy for overall course speed. As shown in Section 3.4, net elevation drop explains only 5% of the cross-course variation in estimated speed effects. The reason is straightforward: downhill can matter, but it is only

one component of course difficulty. Gross climbing, altitude, surface, temperature, course layout, and other features also shape performance, and many of these factors are not well captured by net elevation change.

As a result, the policy misses both fast and slow courses. Some steep downhill races, such as REVEL Big Bear and REVEL Mt. Charleston, do receive substantial penalties, and in those cases the rule moves in the right direction. But many fast courses have little net drop and therefore escape penalty entirely. BQ.2 Spring Chance, for example, has only 80 feet of net drop but an estimated speed advantage of 16 minutes relative to Boston. Under the B.A.A. rule, it would receive no penalty. The converse problem also arises: some courses with large net drops are not especially fast once gross climbing and other features are taken into account. St. George and Big Cottonwood, for instance, have moderate speed advantages, so a flat 5-minute penalty may overcorrect. More generally, any policy based on a single observable proxy will leave much of the relevant venue heterogeneity untouched.

6.2 Course-Specific Qualifying Handicaps

A more coherent alternative is to adjust qualifying standards using the estimated course speed index itself. The logic is simple: the qualifying time at each course should be tightened or loosened by the amount that the course is estimated to differ from Boston. If REVEL Big Bear is 16 minutes faster than Boston, then the comparable qualifying standard for an M18–34 runner would be about 2:44 rather than 3:00. If a course is 5 minutes slower than Boston, the corresponding standard would be 3:05. This approach directly targets the underlying object of interest—how fast a course actually is—rather than relying on a single imperfect proxy.

In practice, the implied adjustments are substantial. Eighty-seven courses in the index are estimated to be more than 5 minutes faster than Boston and would therefore warrant meaningful tightening under a venue-adjusted system. At the same time, many slow courses would warrant looser standards. This symmetry is important. The current debate often

frames the problem as one of a few runners gaining an unfair advantage on extreme downhill courses. But the same logic implies that runners on slow courses are already being penalized by a common standard that ignores large differences in venue difficulty. A separate but related question is whether the age- and gender-specific adjustments in the current BQ system equalize qualification rates across demographic groups; Appendix C examines this question and finds that younger runners and men face somewhat more favorable effective thresholds.

6.3 Counterfactual Qualifying Fields

What would happen if qualifying standards were adjusted using the course speed index? Figure 7 summarizes the answer. Panel A compares the B.A.A.'s net-elevation penalties with the AKM-based handicaps. The mismatch is substantial: many courses with large estimated speed advantages receive no penalty under the B.A.A. rule, while some penalized downhill courses appear to be overadjusted relative to their actual speed.

Panel B reports the resulting counterfactual qualifying field. Under FE-adjusted standards, 13,664 runners who qualify under the current system would lose qualification, all from fast courses, while 8,412 runners who currently miss qualification would gain it, all from slow courses. The net effect is a reduction of 5,252 qualifiers and a turnover rate of 34%. This reallocation is the key policy result. Venue-adjusted standards do not merely exclude runners from a handful of extreme downhill races; they substantially rebalance the field by tightening the bar where courses are unusually fast and loosening it where courses are unusually slow.

The detailed course-by-course winners and losers under FE-adjusted qualifying are shown in Appendix Figure 23.

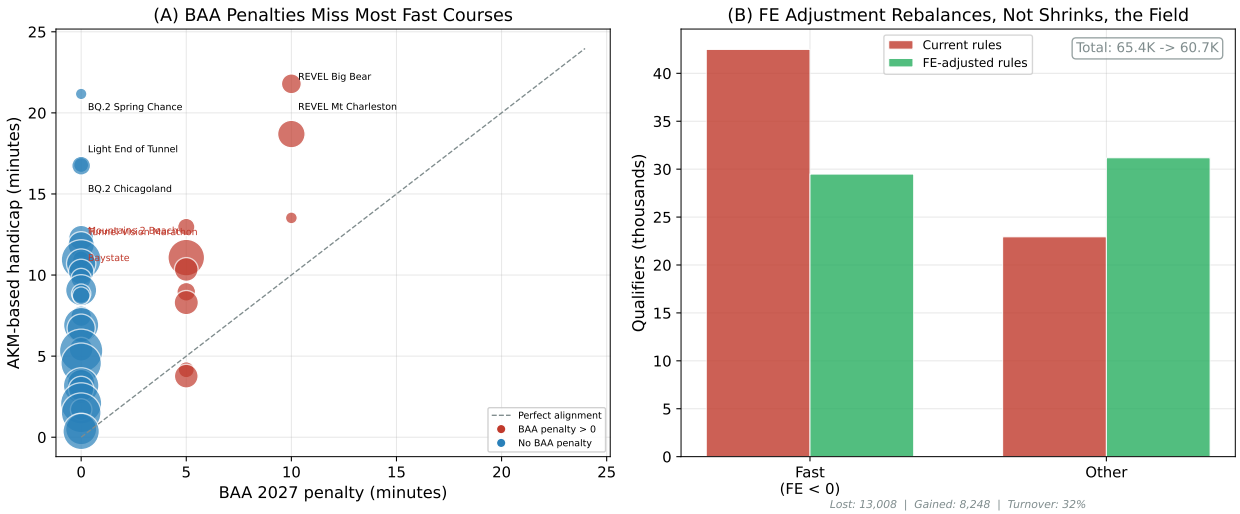


Figure 7: Proxy corrections miss the problem; AKM handicaps rebalance the field. Panel A: BAA 2027 net-elevation penalties vs. AKM-based handicaps. Many fast courses receive zero BAA penalty despite large AKM handicaps. Panel B: qualifiers by course type under current vs. FE-adjusted rules.

6.4 Limitations

Three limitations are worth noting. First, the analysis does not observe intent. I can identify that runners near the margin disproportionately qualify from faster courses, but I cannot distinguish runners who deliberately sought out those venues from runners who chose them for unrelated reasons, such as geography, scheduling, or scenery, and then benefited from the speed advantage. The decompositions therefore speak to the aggregate consequences of course heterogeneity, not to the motives behind any particular runner's choice.

Second, the AKM framework imposes an additive structure on course effects. In the model, a course that is faster than Boston shifts predicted log time by the same amount for all runners after accounting for age and gender. If some course features interact systematically with runner type—for example, if downhill terrain benefits faster runners more than slower runners, or if heat affects weaker runners disproportionately—the estimated course effect will average over those interactions rather than fully capture them. The portability exercises suggest that this approximation is useful in practice, but it is still a simplification.

Third, the counterfactuals are partial-equilibrium exercises. They show how qualification

outcomes would change if standards were adjusted using the estimated course index while holding the underlying course market fixed. In practice, a policy change of that sort could alter demand across races, reduce the premium attached to purpose-built fast courses, and induce organizers and runners to adjust in ways that would change the future distribution of courses and entrants. The counterfactuals should therefore be interpreted as describing the immediate reallocation implied by the current market structure, not as a full equilibrium forecast.

7 Conclusion

The central problem is not simply that some runners seek out downhill marathons. It is that a common qualifying threshold is applied across courses that differ substantially in speed, so the rule confounds underlying ability with venue. As the threshold tightens, that confounding becomes more important. This paper shows that the distortion operates through three linked margins. First, course heterogeneity creates a basic comparability problem: estimated course effects span more than 100 minutes across road marathons. Second, tighter standards amplify those differences by shrinking the set of courses on which marginal runners can qualify and increasing the share of qualifiers drawn from fast venues—a shift driven primarily by reallocation across course types, with within-type runner composition contributing little. Third, runners near the qualifying threshold are disproportionately represented on fast courses, and this excess grows over time; bunching and course-choice evidence indicates that this reflects strategic venue selection rather than differential effort conditional on course choice. The resulting rise in fast-course qualification therefore reflects the interaction of tighter standards with a heterogeneous and evolving course market, amplified by selective sorting at the margin.

Methodologically, the paper uses the AKM framework in reverse. Rather than treating sorting as a nuisance to be differenced away, I use estimated runner and course effects

to measure the extent and consequences of sorting itself. The runner fixed effects provide individual-level counterfactual performance across venues, which makes it possible to separate the contributions of market structure, threshold tightening, and individual course choice in a way that standard bunching approaches cannot. More broadly, the analysis applies wherever candidates are evaluated in environments that differ in difficulty but are judged against a common threshold. In such settings, a common rule will increasingly screen on place as well as person, and ad hoc corrections aimed at a few extreme venues will not solve the underlying problem. A more coherent approach is to adjust standards to the environments in which performance is measured.

References

Abowd, John M., Francis Kramarz, and David N. Margolis. 1999. "High Wage Workers and High Wage Firms." *Econometrica*, 67(2): 251–333.

Allen, Eric J., Patricia M. Dechow, Devin G. Pope, and George Wu. 2017. "Reference-Dependent Preferences: Evidence from Marathon Runners." *Management Science*, 63(6): 1657–1672.

Andrews, Martyn J., Leonard Gill, Thorsten Schank, and Richard Upward. 2008. "High Wage Workers and High Wage Firms." *Economica*, 75(299): 417–442.

Baker, George P. 1992. "Incentive Contracts and Performance Measurement." *Journal of Political Economy*, 100(3): 598–614.

Boston Athletic Association. 2025. *2027 Boston Marathon Qualifying Standards and Course Adjustment Policy.* BAA. `https://www.baa.org/races/boston-marathon/qualify`.

Campbell, Donald T. 1979. "Assessing the Impact of Planned Social Change." *Evaluation and Program Planning*, 2(1): 67–90.

Card, David, Jörg Heining, and Patrick Kline. 2013. "Workplace Heterogeneity and the Rise of West German Wage Inequality." *Quarterly Journal of Economics*, 128(3): 967–1015.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014*a*. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review*, 104(9): 2593–2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014*b*. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review*, 104(9): 2633–2679.

Correia, Sergio. 2017. "Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator." *Working Paper*. `http://scorreia.com/research/hdfe.pdf`.

Goodhart, Charles A.E. 1984. "Problems of Monetary Management: The U.K. Experience." *Monetary Theory and Practice*, 91–121.

Holmström, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization*, 7: 24–52.

Jacob, Brian A., and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics*, 118(3): 843–877.

Kleven, Henrik J., Martin B. Knudsen, Claus Thustrup Kreiner, Soren Pedersen, and Emmanuel Saez. 2011. "Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark." *Econometrica*, 79(3): 651–692.

Lazear, Edward P. 2006. "Speeding, Tax Fraud, and Teaching to the Test." *Quarterly Journal of Economics*, 121(3): 1029–1061.

MarathonGuide.com. 2024. "MarathonGuide.com: Race Results Database." `https://www.marathonguide.com`.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics*, 142(2): 698–714.

Morris, Carl N. 1983. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association*, 78(381): 47–55.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics*, 92(2): 263–283.

Saez, Emmanuel. 2010. "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy*, 2(3): 180–212.

Appendix A Data Construction

Section 2 describes the primary data sources and estimation sample. This appendix provides additional detail on the entity-resolution procedure and the selection of the race-time variable.

A.1 Entity Resolution

I link runner records using a nine-pass entity-resolution procedure. Pass 1 groups observations by exact matches on cleaned name, gender, and imputed year of birth, and links 91.5% of records. Passes 2–9 then progressively relax the matching criteria. Pass 2 imputes missing years of birth using the within-name modal value. Passes 3–5 allow common nicknames and one-year discrepancies in year of birth. Passes 6–9 address abbreviations and other name variants. As a robustness check, I re-estimate the main course effects using only high-confidence exact matches. The resulting estimates have a Spearman correlation of 0.999 with the baseline estimates.

A.2 Time Variable Selection

Web-scraped race results sometimes contain misaligned columns, so a variable labeled "time" may correspond to gun time, chip time, or an unrelated field. For each race-year, I therefore select the time variable that minimizes within-runner mean squared prediction error, using the runner's times in other races as the prediction benchmark. When both gun and chip times are available, this procedure selects chip time—the faster measure—in 91.5% of cases. The average within-runner difference between gun and chip time is about one minute, which is small relative to the interquartile range of estimated course effects (11 minutes).

Appendix B Robustness

I estimate nine alternative specifications. Three progressively exclude the noisiest runners, defined as the top 1%, 5%, and 10% by within-runner mean squared AKM residual. The remaining six alter the sample by excluding trail races, restricting to high-confidence entity-resolution matches, limiting the sample to post-2010 observations, requiring at least 3 or 5 appearances per runner, or restricting to races with at least 500 observations. All nine specifications yield Spearman rank correlations above 0.98 with the baseline estimates. The race-500 specification is nearly identical to the baseline (Spearman 0.9999). The largest deviation arises in the post-2010 sample (Spearman 0.977), consistent with some courses changing character over time.

Robustness to Strategic-Effort Contamination. A concern for the effort decomposition in Section 5 is that the runner fixed effects may themselves be contaminated by strategic effort near the BQ threshold. If runners systematically exert more effort when close to qualifying, their estimated baseline performance could be biased upward, which would in turn bias the effort residuals. I address this concern in two ways.

First, I re-estimate the AKM model after excluding runners within 5 or 10 minutes of the BQ cutoff. I also estimate the model on a sample restricted to inframarginal runners, defined as those whose predicted baseline performance is more than 5 minutes slower than the BQ threshold on a neutral course. In this sample, proximity to the threshold cannot affect effort. In all three cases, the estimated course effects are substantively unchanged relative to the baseline, indicating that the course-speed index is not driven by strategic effort.

Second, I re-estimate the full AKM specification after dropping all observations within ± 5 minutes of the BQ cutoff. This produces "clean" runner fixed effects identified only from observations for which runners had no immediate BQ incentive to exert differential effort. These fixed effects are highly correlated with the baseline estimates, confirming that the individual counterfactuals used in the effort decomposition are robust.

Figure 8 plots estimated course effects for 10 key races across all robustness specifications.

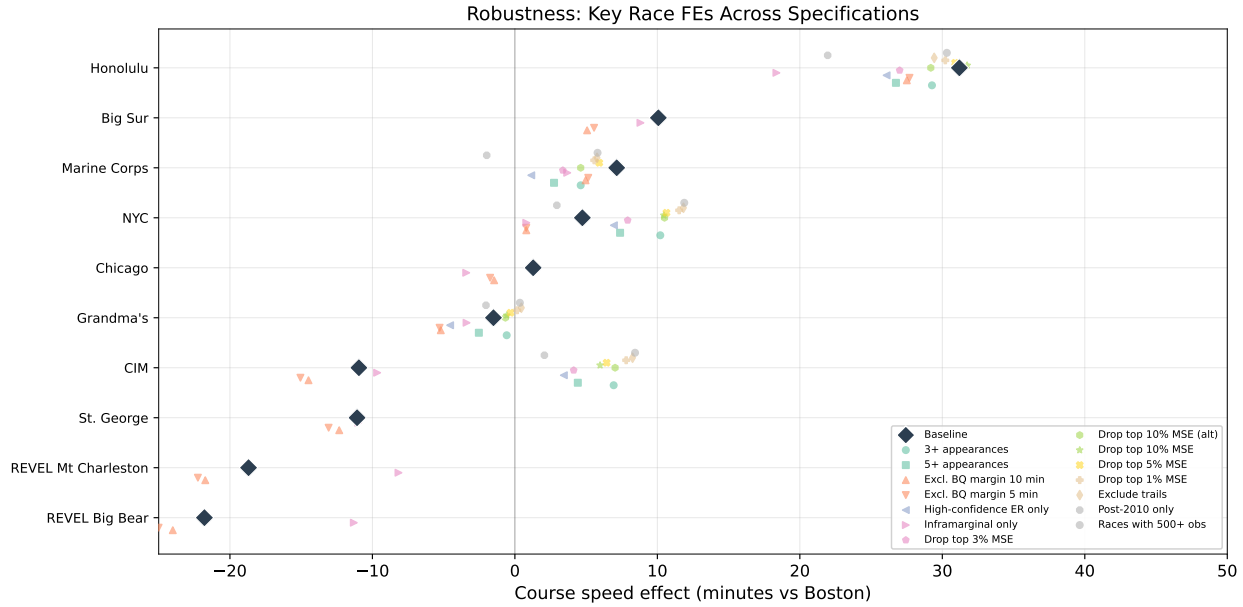


Figure 8: Course speed effects for 10 key races across robustness specifications. Diamond: baseline. Small markers: alternative specifications.

Appendix C Age and Gender Equity

Current BQ standards vary by age group and gender. I examine whether these standards equalize qualification rates across groups. Under course-speed-adjusted standards, younger age groups have slightly more favorable qualification rates than older age groups. Women also have lower qualification rates than men in most age groups, despite the 30-minute allowance built into the current BQ standards. Figures 9 and 10 show these patterns.

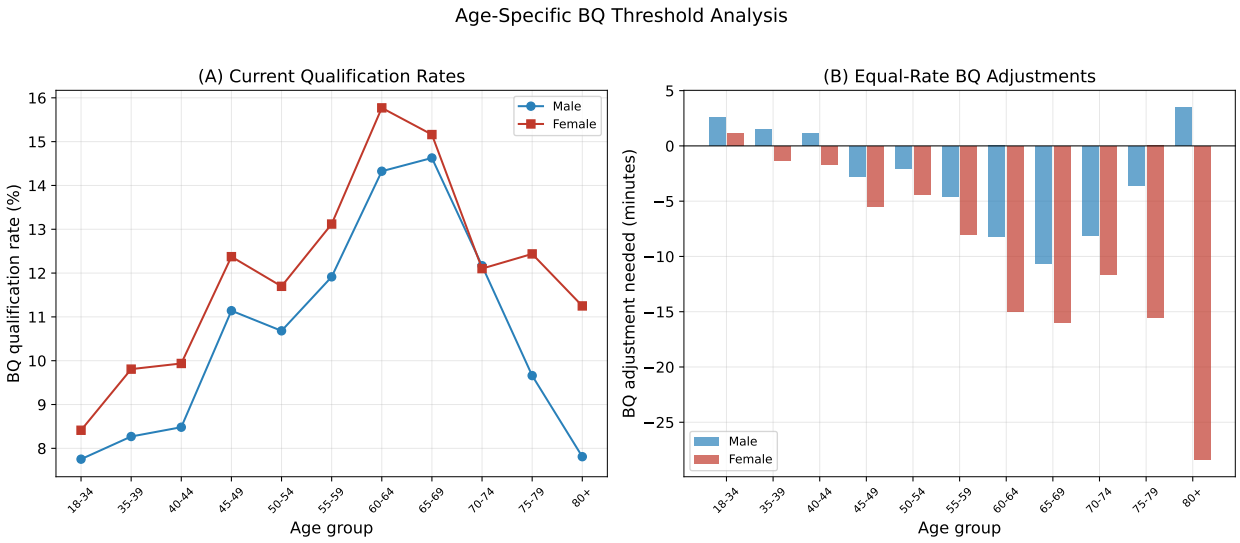


Figure 9: BQ qualification rates by age group under current and equalized standards.

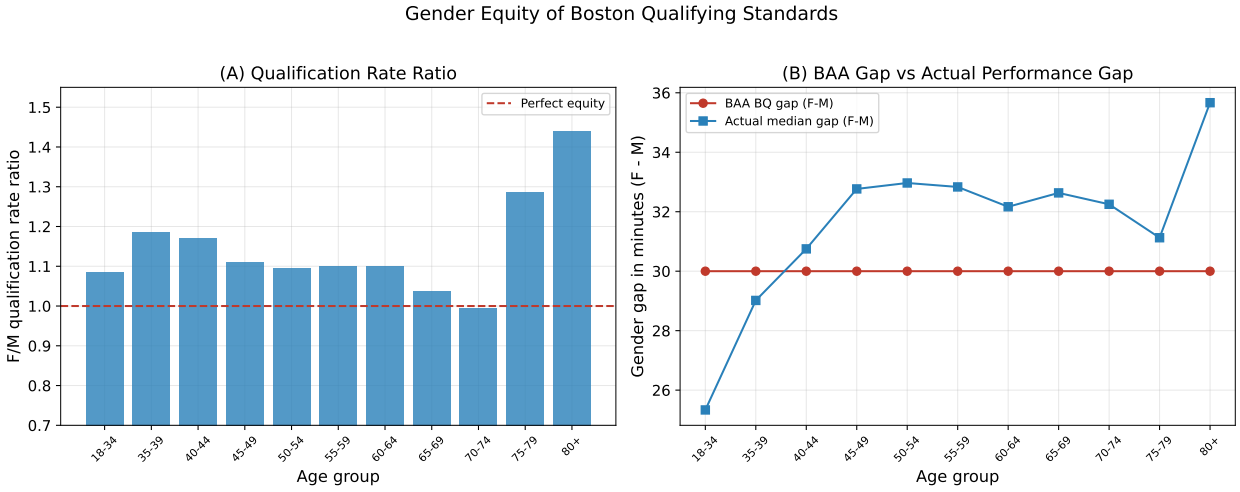


Figure 10: BQ qualification rates by gender and age group.

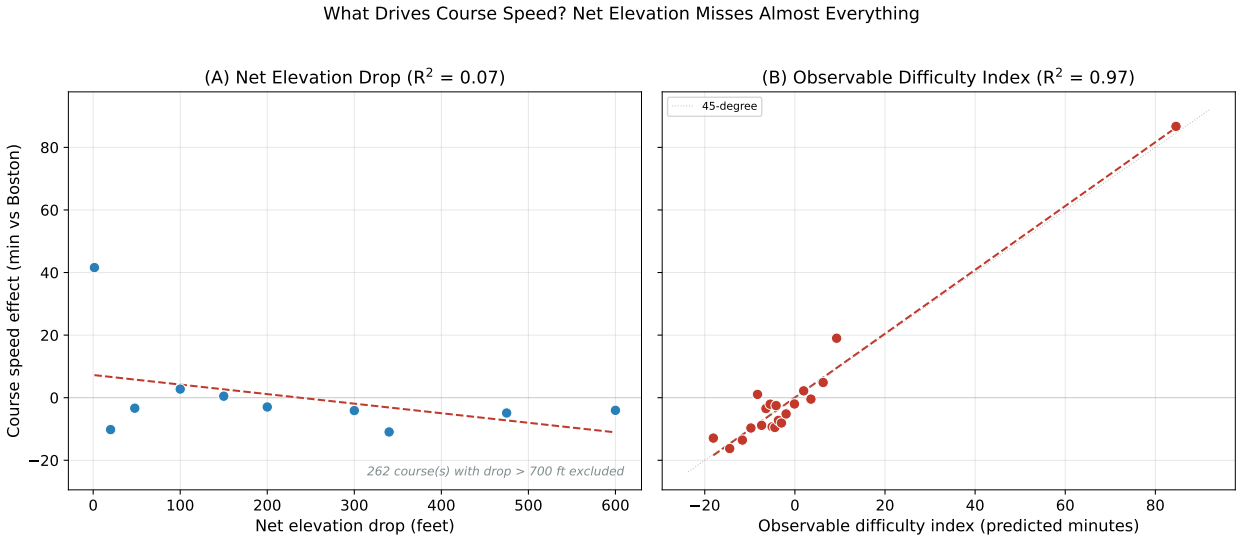


Figure 11: Course speed effects and observable course characteristics. Panel A plots estimated course speed effects against net elevation drop ($R^2 = 0.05$). Panel B plots estimated course speed effects against the observable difficulty index ($R^2 = 0.85$). Each panel shows vigintile binscatters.

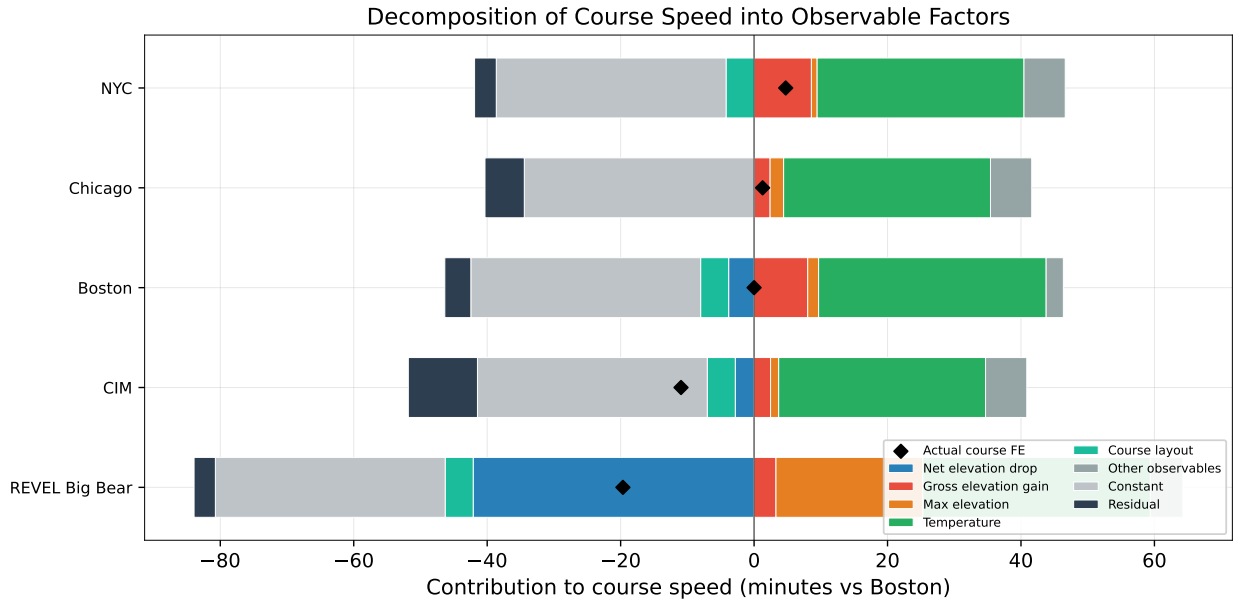


Figure 12: Observable-characteristics decomposition for six representative marathons. Each bar shows the contribution of observable course characteristics to the estimated speed effect, in minutes relative to Boston. The gray residual is the portion not explained by observables.

Table 6: Course fixed effects and net elevation in predicting Boston slowdown

	(1) Full FE	(2) Elevation	(3) Both	(4) Marginal FE	(5) Both
Course FE (full sample)	-0.407^{***}		-0.212^{***}		
	(0.087)		(0.053)		
Course FE (marginal)				-0.545^{***}	-0.314^{***}
				(0.112)	(0.077)
Elev. drop (1000 ft)		0.0146^{***}	0.0105^{***}		0.00972^{***}
		(0.0011)	(0.0014)		(0.0013)
N	35,695	35,762	35,695	35,733	35,733
R^2	0.023	0.029	0.033	0.026	0.034

Notes: The dependent variable is log Boston finish time minus log qualifying finish time. "Full" course FE denotes the shrunk AKM estimate from the full sample. "Marginal" course FE denotes the corresponding estimate from the near-BQ sample. Net elevation drop is measured in thousands of feet. Standard errors are in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

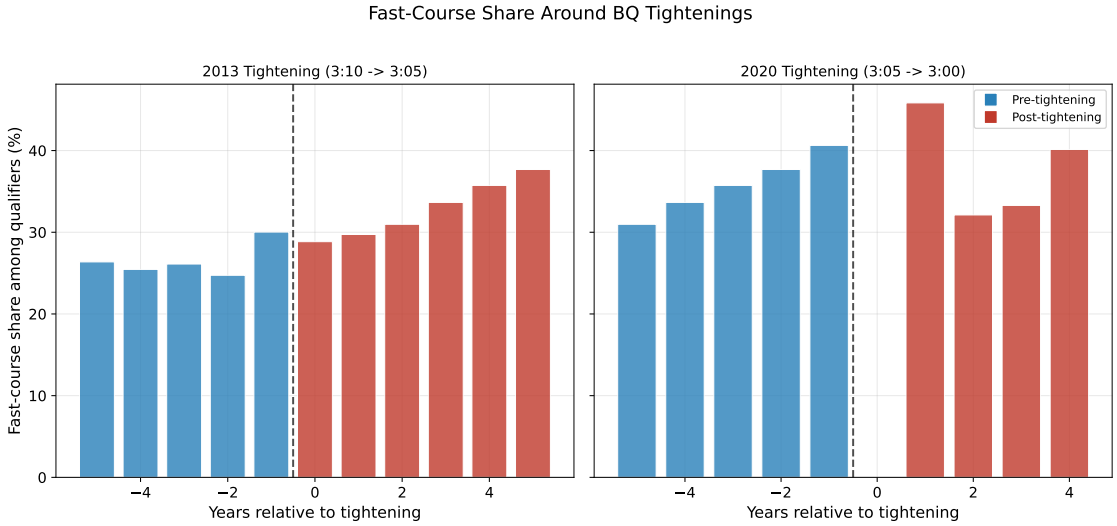


Figure 13: Fast-course share among BQ qualifiers around qualifying-standard tightenings. Event-study evidence for the 2013 and 2020 BQ tightenings.

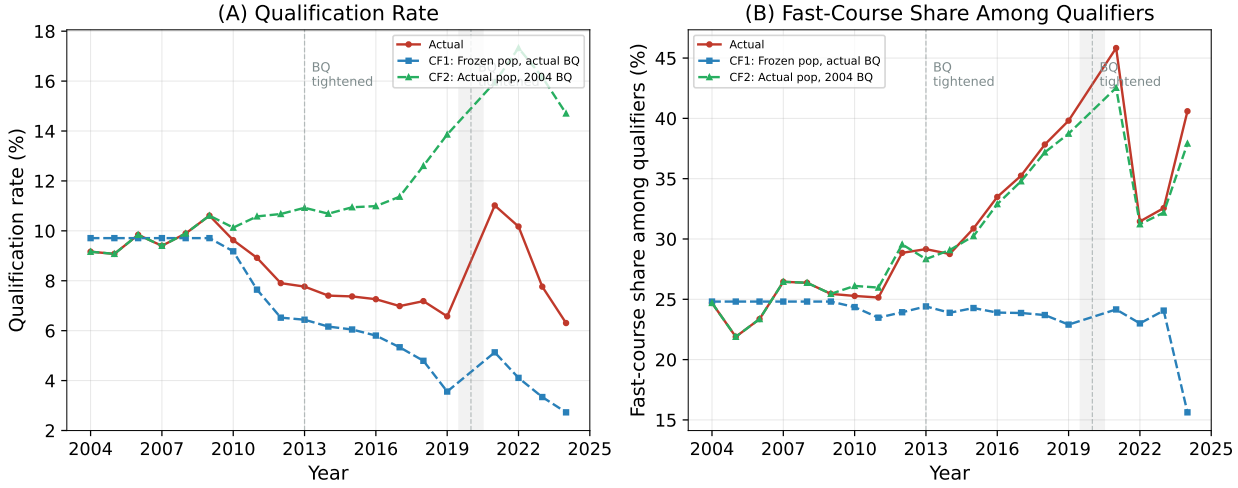


Figure 14: Frozen-population counterfactual decomposition. The figure plots the actual fast-course qualifier share (solid; fast courses are more than 5 minutes faster than Boston), CF1 with the 2004–2009 runner population held fixed and actual BQ standards (dashed), and CF2 with the actual runner population and 2004 BQ standards held fixed (dotted).

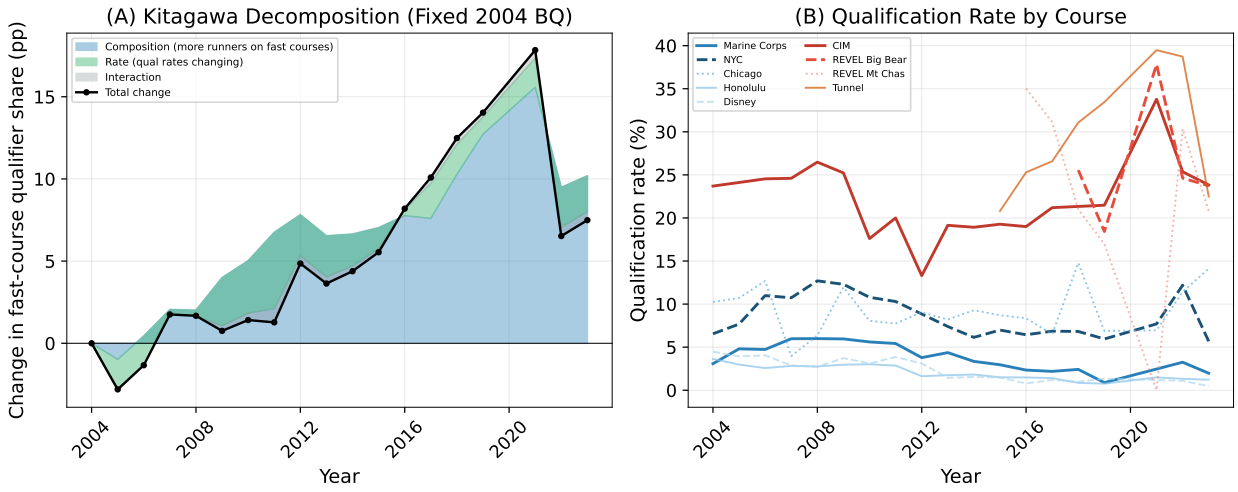


Figure 15: Kitagawa decomposition of changes in the fast-course qualifier share (fast courses are more than 5 minutes faster than Boston). Panel A decomposes the change into composition, rate, and interaction components. Panel B plots qualifier shares over time for selected fast and slow courses.

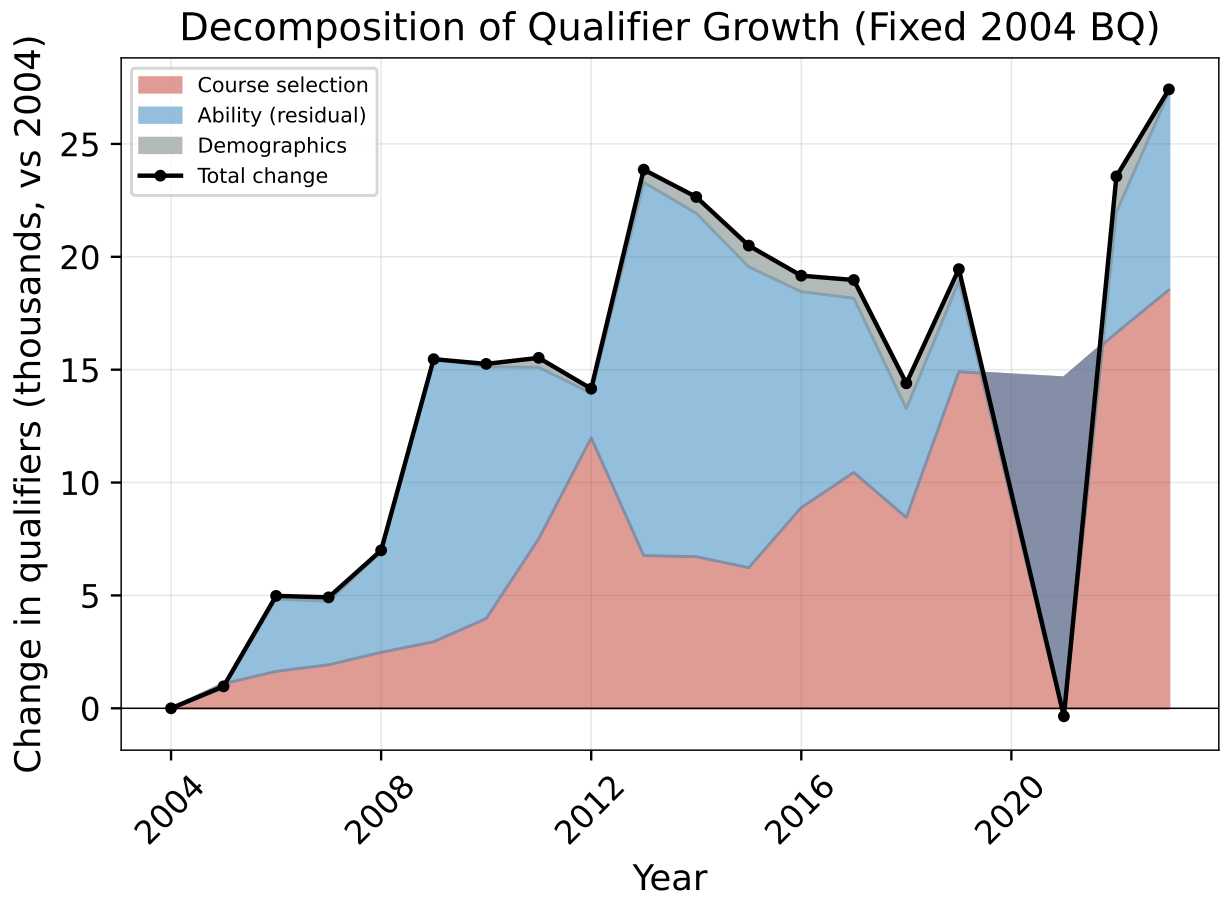


Figure 16: Sequential reweighting decomposition of qualifier growth relative to the 2004 baseline.

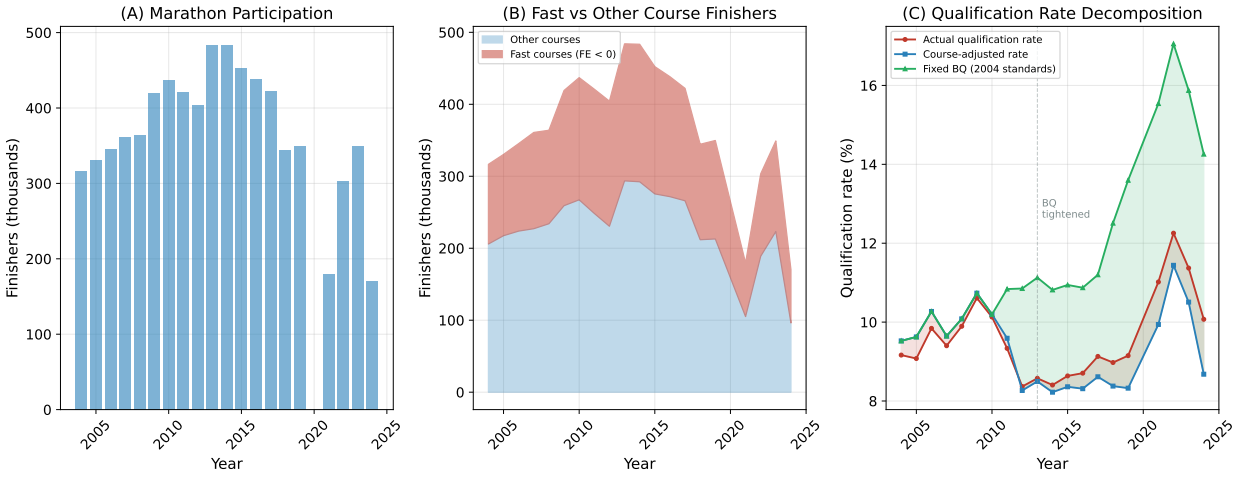


Figure 17: Decomposition of the qualification rate. Panel A plots total U.S. marathon finishers. Panel B plots finishers at fast courses (FE < 0) and other courses. Panel C plots actual, course-adjusted, and fixed-BQ qualification rates.

Appendix D Additional Measurement Figures

Appendix E Additional Mechanical Decomposition Figures

Appendix F Additional Behavioral Response Figures

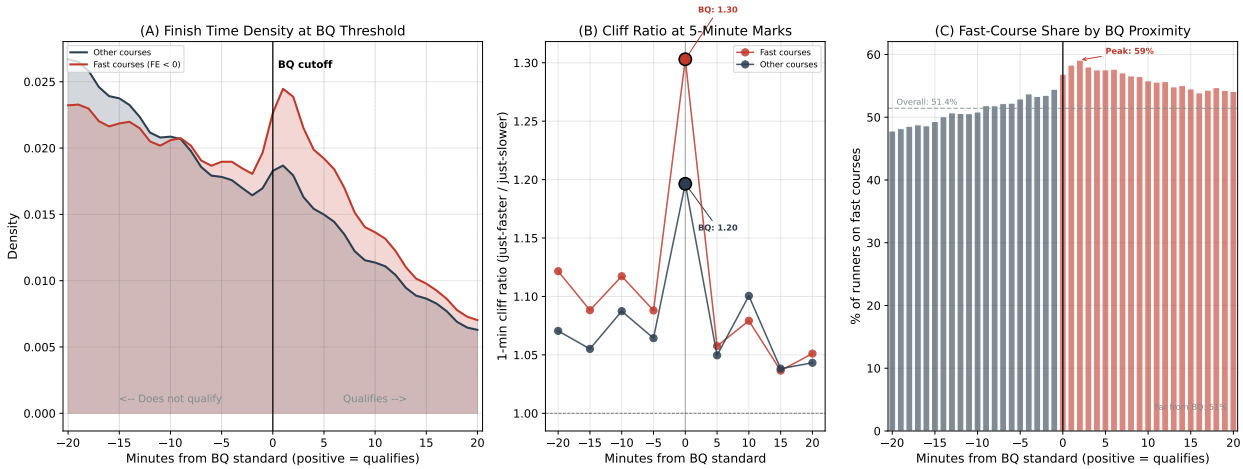


Figure 18: BQ-centered bunching pooled across all age-gender groups. Panel A plots smoothed finish-time densities for fast courses (FE < 0) and other courses. Panel B plots the cliff ratio at each 5-minute mark relative to the BQ threshold. Panel C plots the fast-course share by finish-margin bin.

Appendix G Policy Counterfactual Details

This figure shows how FE-adjusted BQ standards reallocate qualifiers across courses relative to the current system.

Appendix H Bunching Evidence

The AKM runner fixed effects, converted into predicted finish times on a neutral course, imply a baseline performance distribution that is smooth through salient round-number

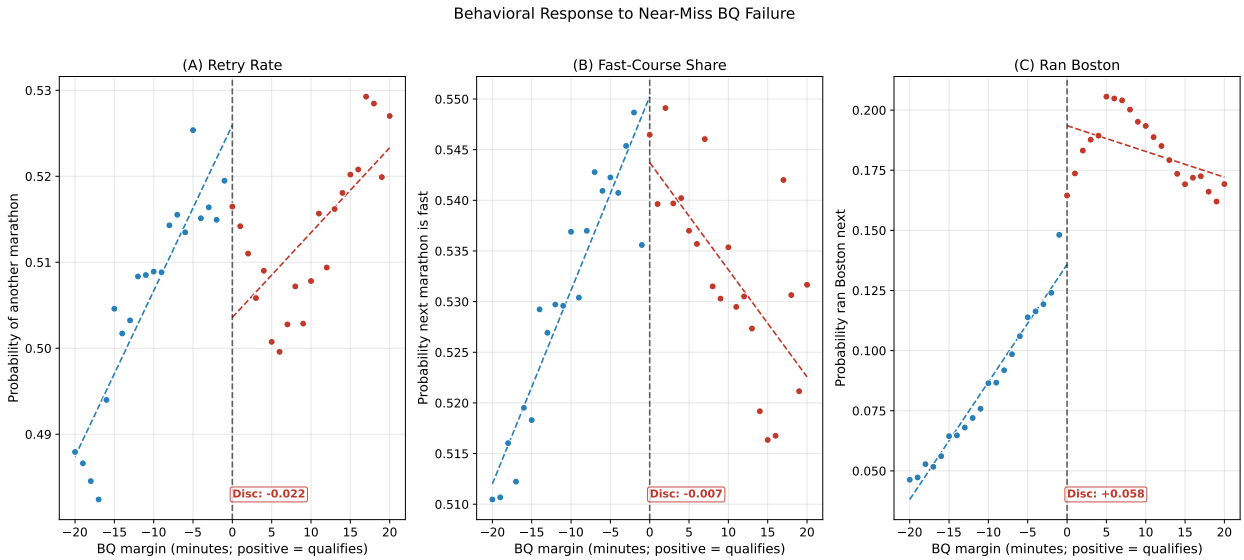


Figure 19: Behavioral responses at the BQ threshold. Panel A plots the probability of re-entry. Panel B plots the fast-course share among re-entrants. Panel C plots the probability of subsequently running Boston.

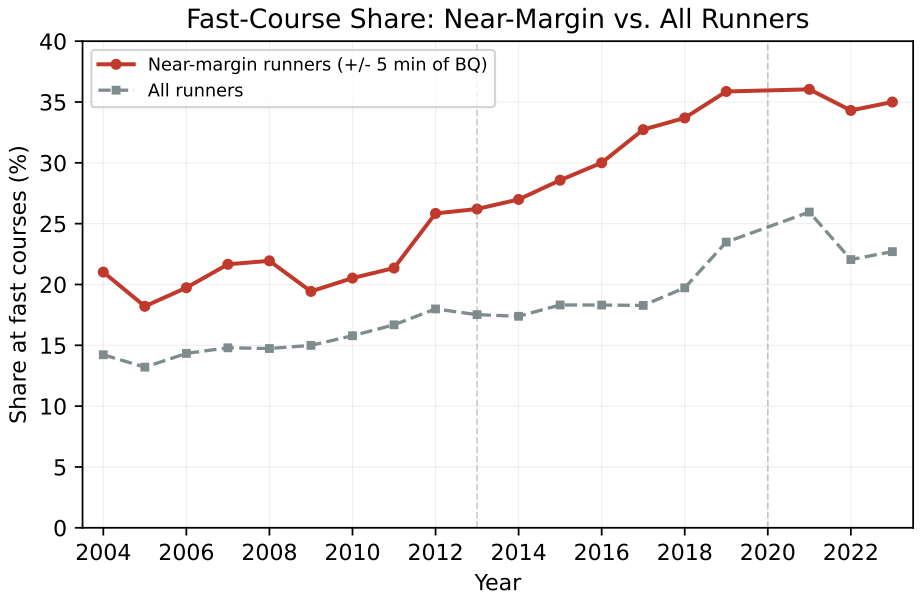


Figure 20: Fast-course share among near-margin runners and all runners over time. Near-margin runners are those whose FE-adjusted Boston-equivalent finish time lies within ±5 minutes of the 2004 BQ standard. Fast courses are those estimated to be more than 5 minutes faster than Boston. The gap between the two series measures excess fast-course representation among runners near the qualifying threshold.

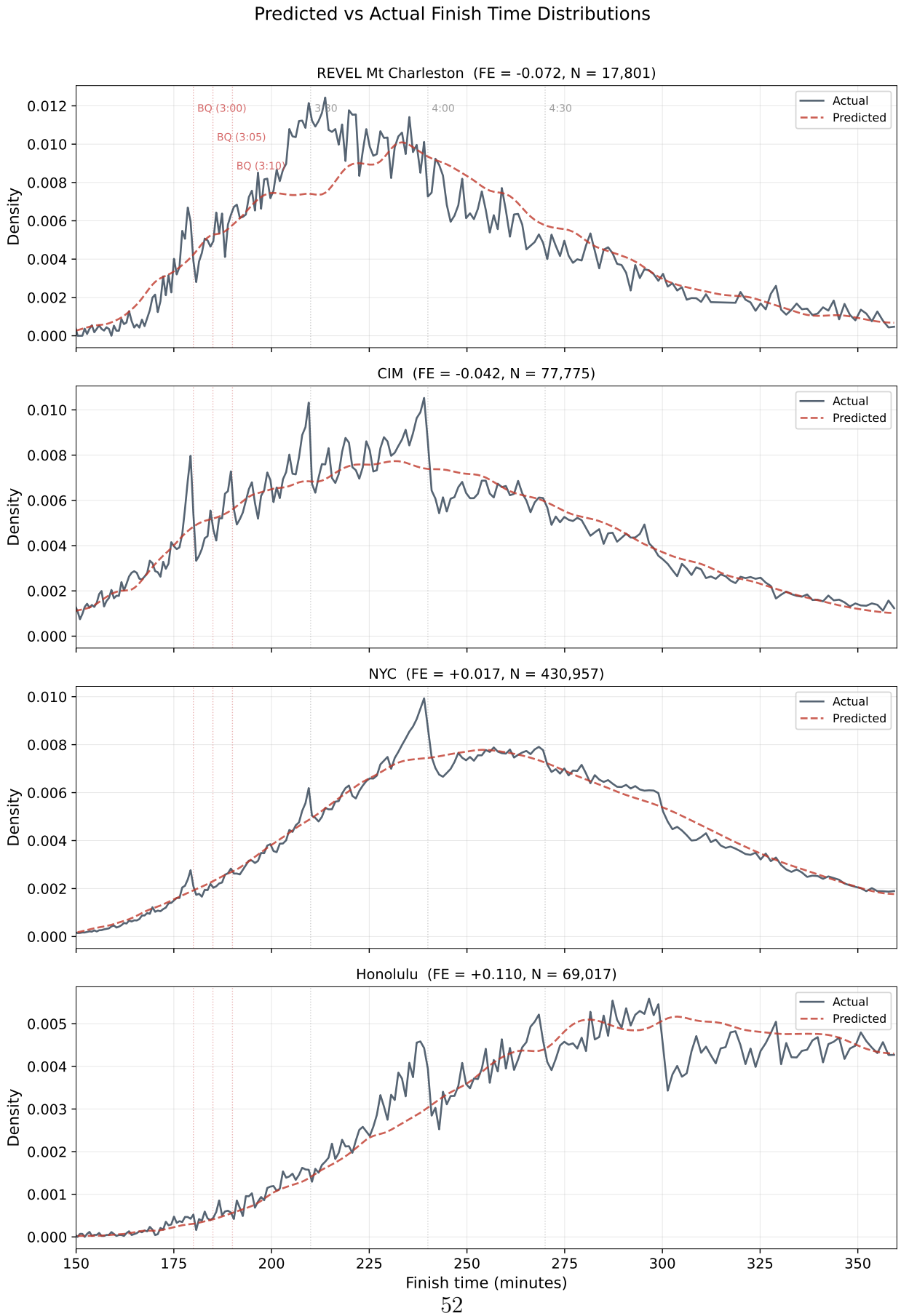


Figure 21: Predicted and actual finish-time distributions for four representative courses. The dashed line shows the predicted distribution, and the solid line shows the actual distribution. Actual finish times exhibit bunching at BQ thresholds and round numbers.

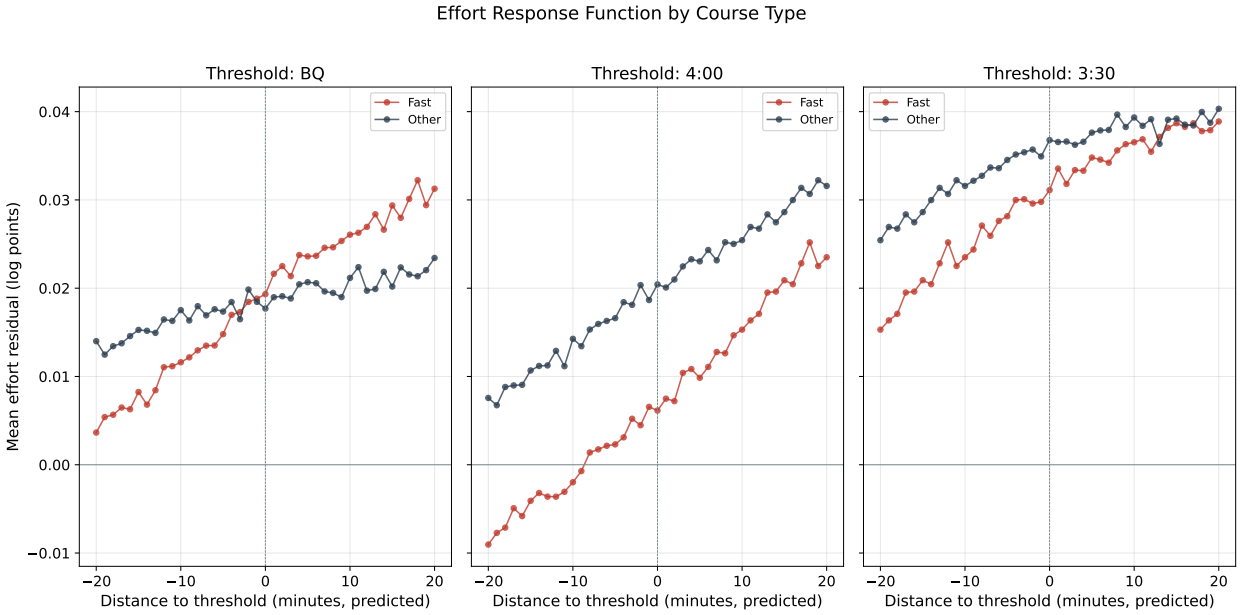


Figure 22: Effort-response function. The figure plots mean log-point effort residuals against distance to the threshold for fast and other courses. At round-number thresholds, the two series largely overlap. At BQ thresholds, they diverge.

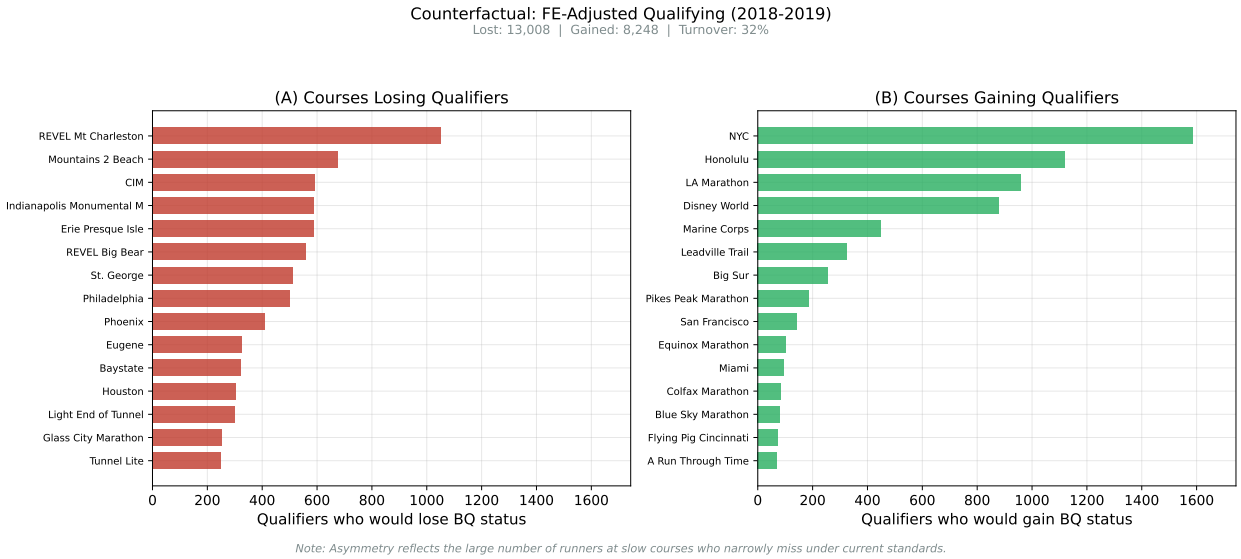


Figure 23: Counterfactual distribution of qualifiers across courses under FE-adjusted BQ standards. Panel A shows courses that lose qualifiers relative to the current system. Panel B shows courses that gain qualifiers.

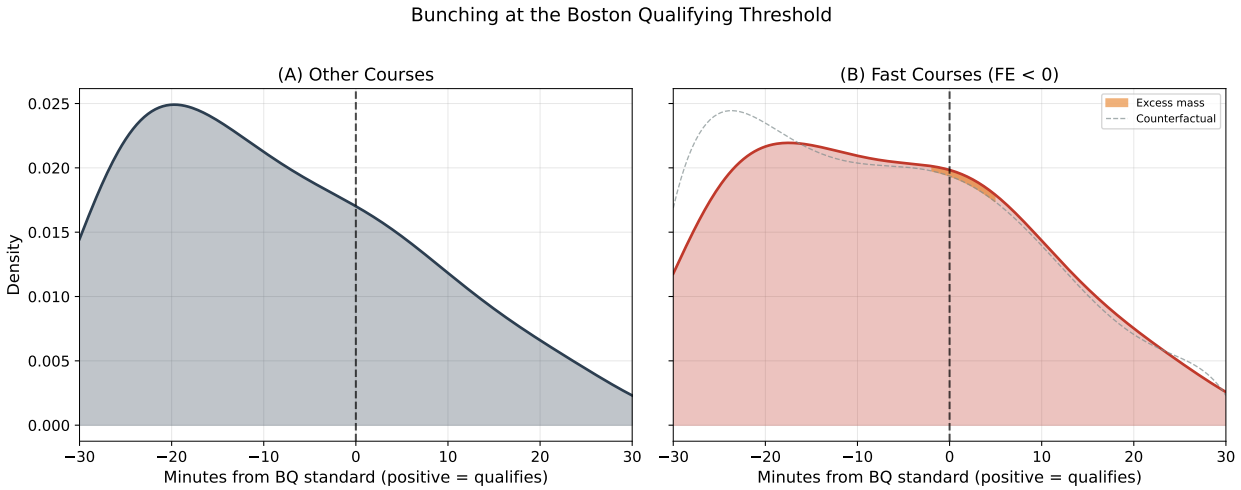


Figure 24: Bunching at the BQ threshold. Panel A shows finish-time density around the cutoff for other courses. Panel B shows the corresponding density for fast courses. The shaded region denotes excess mass relative to a polynomial counterfactual. Positive values on the x-axis indicate qualification.

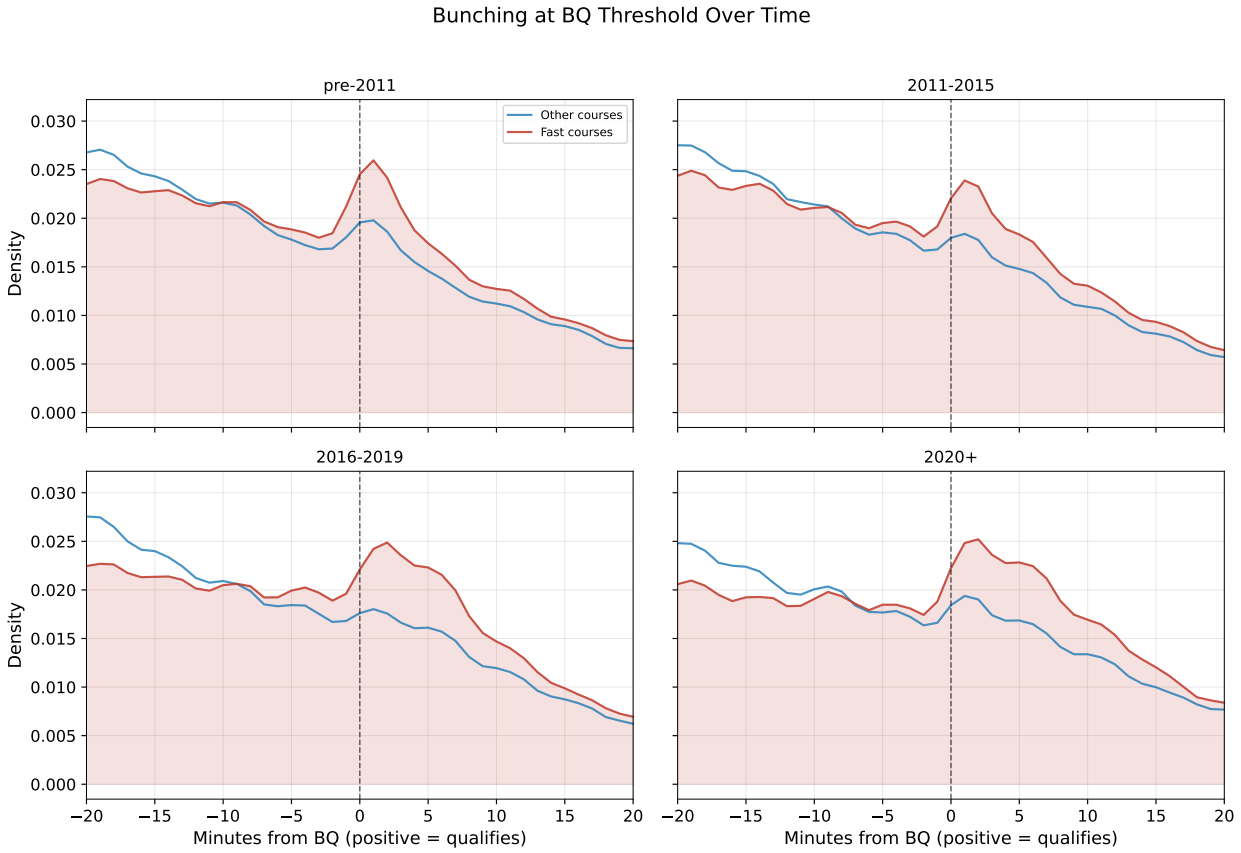


Figure 25: BQ-margin density by era for fast courses (FE < 0) and other courses. Excess bunching on fast courses increases from the pre-2011 period to the 2020+ period.

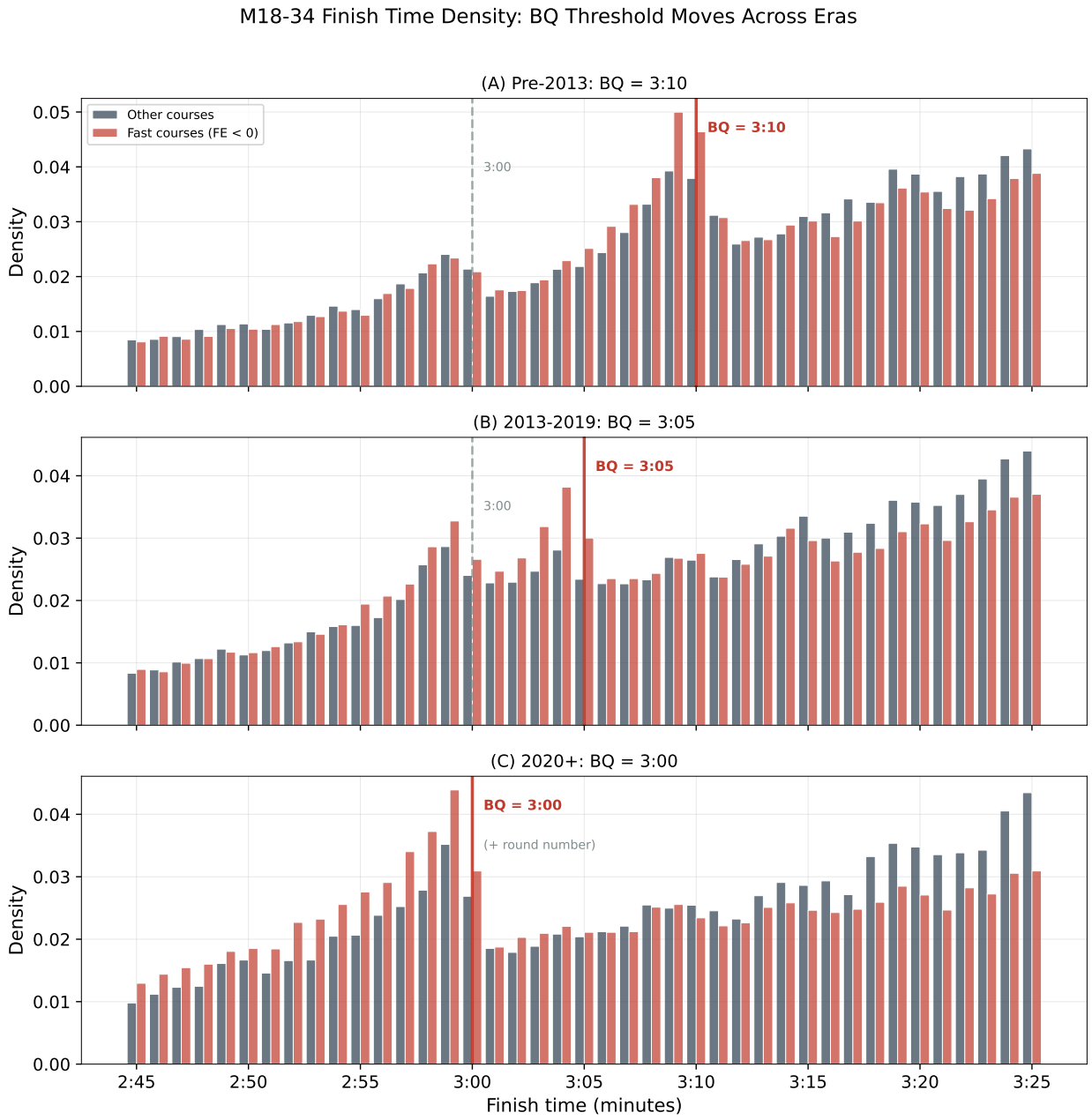


Figure 26: Finish-time density for men ages 18–34 by era. The bunching spike follows the BQ threshold (3:10 → 3:05 → 3:00) rather than fixed round numbers. When 3:10 ceases to be the BQ standard after 2013, the corresponding spike disappears.

thresholds (Figure 27, Panel A). Actual finish times, by contrast, display pronounced spikes at half-hour and hour marks (Panel B). This contrast indicates that the bunching documented in Section 5 reflects behavior rather than irregularities in the underlying ability distribution.

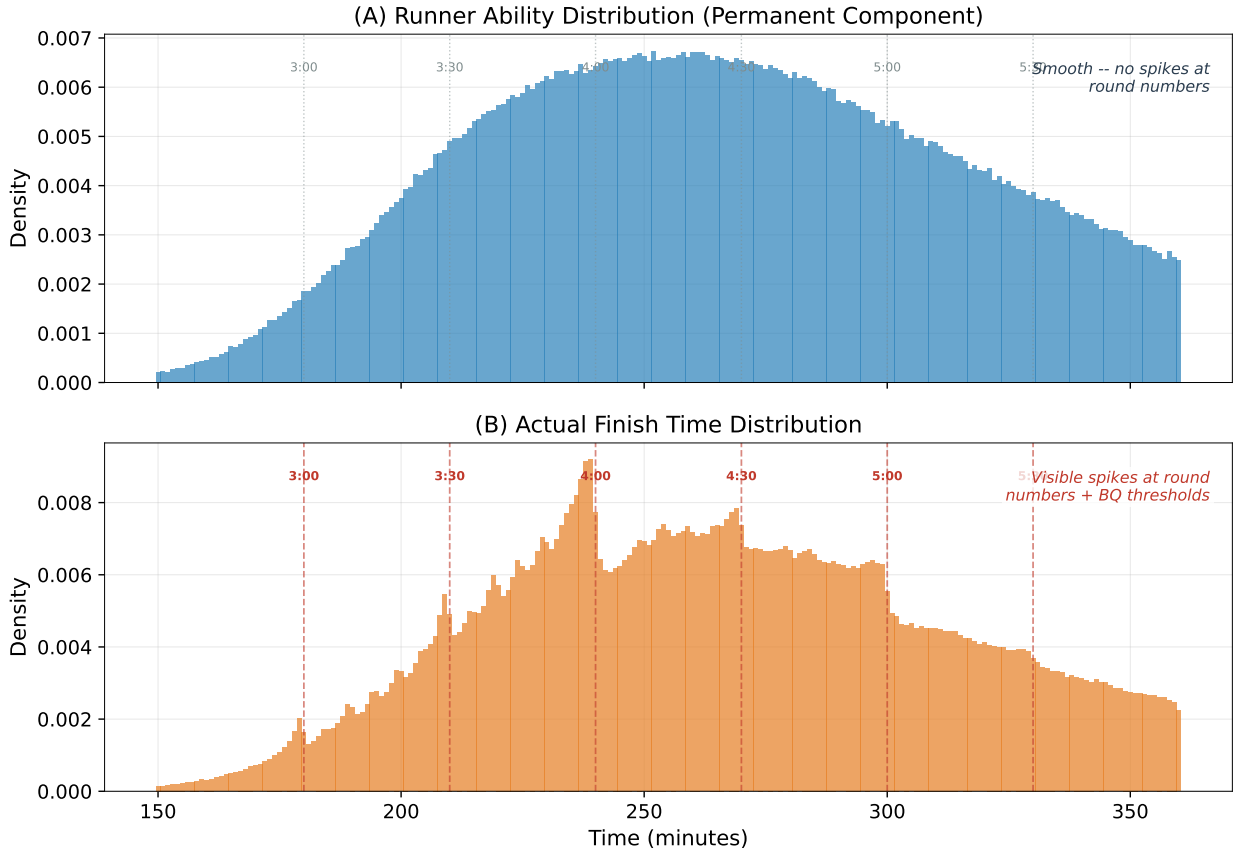


Figure 27: Baseline runner performance and actual finish times. Panel A plots the distribution of predicted neutral-course finish times. Panel B plots the distribution of actual finish times. The baseline distribution is smooth, whereas actual finish times bunch at salient thresholds.

A further concern is that sorting into fast courses may reflect generic round-number pacing rather than BQ-specific incentives. I assess this by examining the 4:00 and 3:30 marks after excluding runners for whom those times are themselves BQ thresholds (Figure 28). At 4:00, the fast-course share is essentially flat across the threshold (39.6% versus 37.9%). At 3:30, it is similarly flat (45.5% versus 45.7%). At the BQ threshold, by contrast, the fast-course share rises from 48.2% to 52.7%. The same pattern appears in the cliff ratios: at

4:00, the fast-course ratio (1.55) is similar to the other-course ratio (1.45), whereas the two diverge at the BQ threshold (1.38 versus 1.11). This pattern indicates that course sorting is specific to the BQ cutoff rather than a generic pacing artifact.

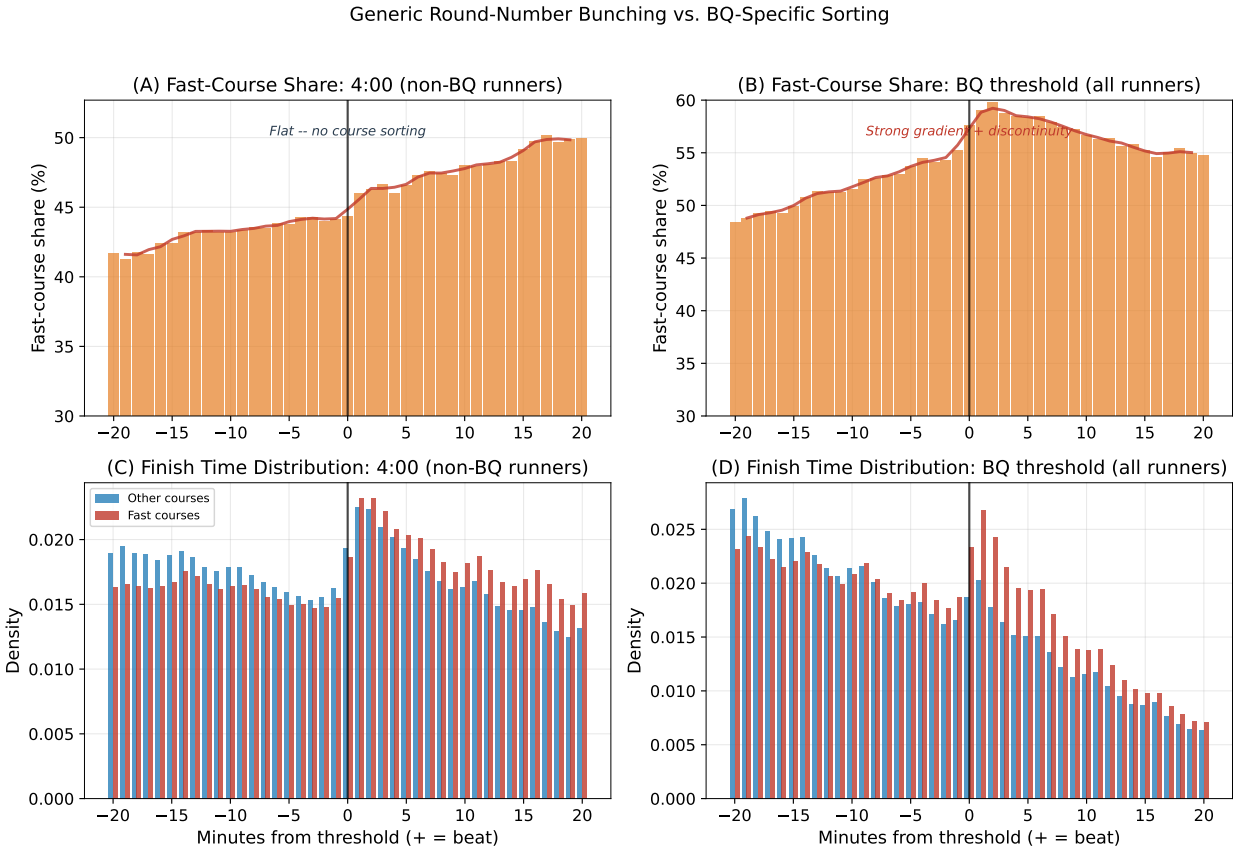


Figure 28: Generic round-number bunching and BQ-specific sorting. The top row examines the 4:00 mark, excluding runners for whom 240 minutes is the BQ threshold. The bottom row examines the BQ threshold. Sorting into fast courses is limited at the generic round number but pronounced at the BQ cutoff.

Appendix I Weather Shocks and Course Switching

A weather-based natural experiment provides additional evidence on strategic course choice. When unusually adverse conditions disrupt a major marathon—for example, an abnormally hot day—some runners who otherwise might have qualified narrowly miss the BQ cutoff. If these runners respond strategically, they should be more likely to re-enter and to choose

a fast course in their next attempt. I identify weather-shock race-years using race-by-year fixed effects: a shock occurs when the race-year effect exceeds the course's historical mean by more than 1.5 standard deviations. I then compare near-BQ misses (0–5 minutes above the cutoff) in shock years with the same group in non-shock years, using runners 15–30 minutes above the cutoff as a placebo group.

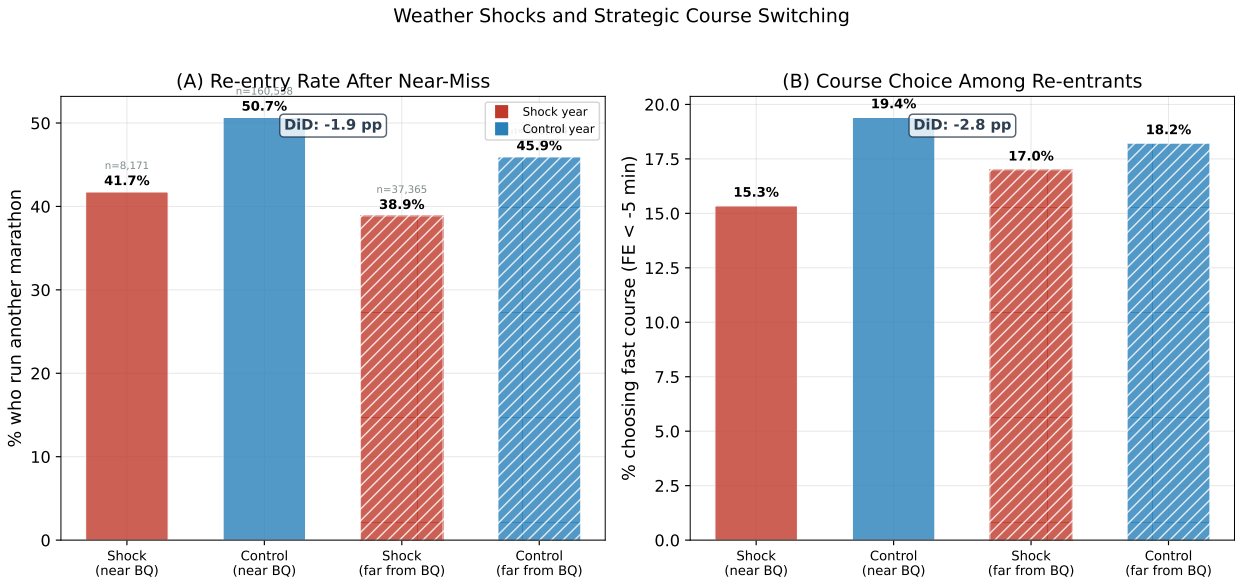


Figure 29: Weather shocks and subsequent race choice. Panel A plots the share of near-BQ and far-from-BQ misses who run another marathon, separately for shock and non-shock race-years. Panel B plots the share choosing a fast course (FE < -5 minutes), conditional on re-entry. Difference-in-differences estimates compare the shock-non-shock gap for near-BQ runners with the corresponding gap for far-from-BQ runners.

Figure 29 shows that weather shocks do not produce differential re-entry: the shock–non-shock gap for near-BQ runners is similar to the corresponding gap for far-from-BQ runners (DiD $= -0.4$ percentage points). Conditional on re-entry, fast-course switching likewise yields a difference-in-differences estimate close to zero. Exogenous shocks to qualifying prospects therefore do not appear to induce runners to switch to faster courses, consistent with substantial inertia in course choice limiting the individual-level sorting response.

Appendix J Predictive Validity of Qualifying Times

A complementary test asks whether FE-adjusting qualifying times improves their ability to predict Boston performance. If the course-speed index captures genuine course-speed differences rather than runner selection, then adjusting qualifying times should do little to change the relationship between qualifying and Boston performance. The speed advantage of fast courses is already embedded in the qualifying time itself.

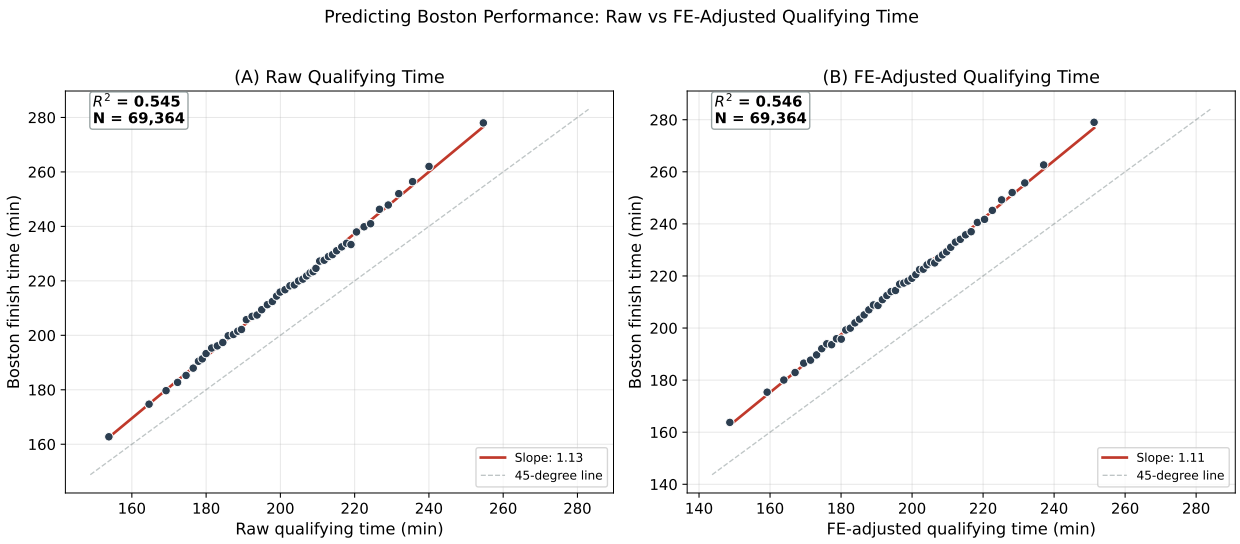


Figure 30: Predicting Boston performance from qualifying time. Panel A plots raw qualifying time against Boston finish time. Panel B plots FE-adjusted qualifying time against Boston finish time. Each point is a quantile-bin mean. The nearly identical R^2 values indicate that FE adjustment does not materially improve prediction.

Figure 30 reports the results. Raw qualifying time predicts Boston finish time with an R^2 of 0.545 and a slope of 1.13. FE-adjusting the qualifying time—removing the estimated course-speed component—barely changes the fit: the R^2 rises to 0.546 and the slope changes to 1.11. This pattern is consistent with the course fixed effect capturing genuine course-speed differences rather than runner sorting. Removing the course component neither improves nor worsens prediction because the speed advantage is real at the qualifying race, but it does not carry over to Boston.