Dynamic Motivation in Goal Pursuit: How Near Misses and Narrow Wins Affect Subsequent Performance

Nicholas C. Owsley[*] George Wu[†] Donovan Rowsey[‡]

September 5, 2026

Abstract

An extensive literature shows that reference points shape effort provision and performance across diverse settings, from competitive running to sales. Yet performance in these settings is rarely one-shot: the runner lines up for another race, the salesperson faces another quarterly target. What ultimately matters for cumulative performance is not only whether someone reaches a threshold but how reaching it, or just missing it, affects motivation in subsequent outings. We use a dataset of 9 million race times from U.S. high school track athletes to examine how "near misses" and "narrow wins" relative to a reference point affect subsequent motivation and performance. We find that boys who narrowly surpass a round-number finishing time are 4.0 percentage points less likely to improve in their next race than boys who narrowly miss one, 3.4 percentage points less likely to improve again that season, and 2.0 percentage points more likely to stop participating in competitive races that season. We find similar effects across multiple middle-distance events and for both girls and boys. These effects are strongest for athletes achieving the threshold for the first time and fade by the next year. The pattern of results is not explained by athletes shifting focus to alternative activities, regression to the mean, changes in risk-taking, or physical exhaustion. Instead, our results are consistent with reference points having a diminished motivational effect once they have been surpassed and inconsistent with standard models of reference dependence.

[*]University of Chicago, Booth School of Business, nowsley@chicagobooth.edu, ORCID: 0000-0001-7898-6947
[†]University of Chicago, Booth School of Business, wu@chicagobooth.edu
[‡]University of Chicago, Booth School of Business; Fetch

1 Introduction

On May 6, 1954, Roger Bannister became the first person to run a mile in under four minutes. After being broken six times in the early 1940s, the mile world record stagnated at 4 minutes and 1.4 seconds between 1945 and 1954. However, after Bannister's historic run, the world record was bested several weeks later, and over the next nine years, 43 different milers broke 4 minutes on more than one hundred occasions. While this pattern of stagnation followed by a deluge of achievement could be idiosyncratic, it could also reflect a psychological effect of the four-minute threshold on individual runners. This possibility has spurred numerous self-help narratives about how beliefs can be self-limiting (Hutchinson, 2018).

In this paper, we investigate the dynamics of goal pursuit at the level of individual athletes. Round numbers, such as a four-minute mile, serve as goals and reference points, dividing outcomes into "gains" and "losses" and motivating individuals to adjust their effort in order to surpass them (Heath, Larrick, & Wu, 1999). For instance, Allen, Dechow, Pope, and Wu (2017b) found that marathon runners modulated their race effort to narrowly surpass round-number finishing times, leading to "bunching" in the distribution of finishing times just beneath round numbers. For example, they found that 50.1% more runners finished a marathon in just under 3 hours (2:59:00 to 2:59:59) than just over 3 hours (3:00:00 to 3:00:59). Similar bunching patterns have been observed in organizational settings, such as around a sales manager's quarterly sales targets (e.g., Pierce, Rees-Jones, and Blank, 2025). However, whereas in organizations thresholds are often tied to material incentives, for most runners the motivation to best a round number threshold is largely psychological and not extrinsic.

In this paper, we leverage a large dataset of U.S. high school running performances to examine how effort and performance change in periods following outcomes close to these reference points. How are people affected by a near miss or a narrow win? How does a runner who narrowly surpasses (misses) a goal fare in their *next* race? Different psychological theories make divergent predictions for how such outcomes shape subsequent motivation. Theories of self-efficacy and "choking" suggest

that narrow wins, relative to near misses, might improve later performance by increasing self-belief and reducing pressure to succeed (Ariely, Gneezy, Loewenstein, & Mazar, 2009; Bandura, 1982; Mesagno & Beckmann, 2017; Yu, 2015). By contrast, the opposite pattern is predicted by the "goal-gradient effect" and accounts of reference-dependent preferences, which show that individuals exert more effort as they move closer to achieving a goal and reduce effort after surpassing it (Cryder, Loewenstein, & Seltman, 2013; Heath, Larrick, & Wu, 1999; Kivetz, Urminsky, & Zheng, 2006; Lukas, 2018).

Recent research has examined how narrow wins and near misses affect future motivation in diverse settings. We review this literature in Section 2. The findings are mixed. For example, Wang, Jones, and Wang (2019), Burdina and Hiller (2021), and Anderson and Green (2018) found negative effects of "narrow wins" relative to "near misses" on subsequent motivation, while Soetevent (2022) and González-Díaz, Palacios-Huerta, and Abuín (2024) found heterogeneous (both positive and negative) or null effects. One illustrative example comes from chess, where González-Díaz et al. (2024) showed that both male and female professional chess players played more often as they approached their personal best ratings, but that the performance of female players *declined* while the performance of men improved as they neared this reference point. They also found mixed effects on effort after players surpassed their personal bests but could not track how these effects evolved over time, leaving open whether such effects persist or decay in the longer run.

We build on this literature by testing how narrow wins and near misses affect subsequent motivation. As in previous investigations, motivation is not directly observed; instead, we observe performance and participation, which we take as proxies for motivation and effort. Our data allow us to conduct highly-powered statistical tests across multiple dimensions of these effects—several measures of performance and persistence—and to examine how the effects evolve over time. The richness of the data also allows us to explore potential mechanisms, including changes in risk-taking, regression to the mean, shifting effort toward alternative activities, and physical exhaustion.

Specifically, our investigation leverages a dataset of 9 million race times for almost one million U.S. high school track athletes. Our data have several compelling features for testing our hypotheses. First, there are clear and stable reference points in this setting. Although we recognize that goals, round numbers, and reference points are distinct constructs, we use the terms interchangeably

throughout the paper. Previous work has shown that round numbers both serve as goals and as reference points in running (Markle, Wu, White, & Sackett, 2018), and that goals can independently serve as reference points (Heath, Larrick, & Wu, 1999) (see also D. Pope and Simonsohn (2011)). We empirically demonstrate the importance of round-number times in this setting by identifying "bunching" in the distribution of finishing times below round numbers (Allen et al., 2017b; Chetty, Friedman, Olsen, & Pistaferri, 2011; Kleven, 2016). Second, outcomes are precisely documented, with times machine-recorded to the 100th of a second. Third, there are largely no external rewards in this setting that could confound the psychological effect of reference point performances. High school track running does not include material compensation, and round numbers generally do not serve as qualifying times for higher-level participation or representation. Finally, our dataset includes results for performances in other competitive events during each individual's high school athletic career. This allows us to test how the effects of near misses and narrow wins change over different time horizons (such as weeks, months, or years), and whether they affect performance and participation in other athletic events.

In organizational settings, this combination of data features is often difficult or impossible to acquire. In these settings, performance data are challenging to precisely measure and reference points are unobserved and heterogeneous across employees (Akerlof & Yellen, 1986; Hansen, 1997; O'Donoghue & Sprenger, 2018). When reference points exist, they are usually linked to external rewards, such as bonuses for exceeding sales targets (Kleven, 2016; Oyer, 1998). Future performance data also typically exist only conditional on past performance: workers are fired, promoted, or otherwise exit the firm—creating selection problems. (Degeorge, Patel, & Zeckhauser, 1999; Oyer, 1998). Finally, precise data on alternative activities are often missing. Our data thus provide a rare opportunity to study how performances relative to "mere reference points" shape subsequent motivation across multiple dimensions of effort and time, and whether individuals shift attention toward or away from alternative activities.

We find that surpassing a reference point has a negative effect on subsequent performance. Athletes who narrowly surpass a round number time are significantly less likely to improve in both the next race and over the remaining track season. High school boys who ran 1600 meters in a seasonal best time of 4:59.00 to 4:59.99 (for short, 4:59) compared to 5:00.00 to 5:00.99 (for short,

5:00) were 4.0 percentage points less likely to improve in their next race, improved 0.36 seconds less, and were 2.0 percentage points less likely to run another race that season (all $p < .001$). The effect is largest the first time an athlete ran faster than the reference point, diminishes over time, and is undetectable by the following season. This effect is highly robust, and also replicates with round number times in other events, specifically the girls 1600-meter race, and boys and girls 800- and 3200-meter races.

What explains this reduction in persistence and performance following a narrow win? We find that regression to the mean, physical exhaustion, and changes in risk-taking cannot account for our results, and while we see some evidence of effort substitution to other events, it is limited and cannot explain our effects. Turning to theoretical accounts, a standard model of reference-dependent preferences does not produce the negative discontinuity in improvement. Our findings are also only partially consistent with the goal gradient effect: we observe the predicted reduction in effort after runners surpass the goal, but not the increase as they approach it. However, a modification to the standard prospect theory model—allowing the reference point to carry less weight after a runner has surpassed it—reproduces our key results. Intuitively, reference points become less motivating once they've been eclipsed. This is consistent with our data showing that the effect is concentrated among runners achieving the reference point for the first time.

This paper makes several contributions. First, we contribute to the literature on the effect of reference-dependent preferences on effort provision (Kőszegi & Rabin, 2006; O'Donoghue & Sprenger, 2018). While most work in this area focuses on the effects of reference points on effort within an activity, we exploit the unique features of our data to demonstrate the effects of performances around a reference point on effort and performance in subsequent, separate activities. We also extend the literature on goals and motivation, building on evidence linking reference-dependent preferences to goal pursuit (Heath, Larrick, & Wu, 1999). While the goal-setting literature has shown that goals are important tools for motivation (Locke & Latham, 1990, 2006), we show that narrowly achieving a round number goal can also hurt future effort, consistent with accounts in which success on a goal licenses reduced striving (Fishbach & Dhar, 2005, 2007). Finally, we contribute to the literature on the dynamic effects of non-linear compensation schemes in organizations (Degeorge et al., 1999; Misra & Nair, 2011; Myers, Myers, & Skinner, 2007; Oyer, 1998). In these settings,

thresholds carry material stakes, making strategic and psychological responses difficult to separate; we show that non-linear *psychological* payoffs alone generate similar dynamic effects on effort.

This paper proceeds as follows. In Section 2, we provide a brief review of the relevant literature. In Section 3, we describe the data and institutional setting. In Section 4, we present the results, and finally, in Section 5, we discuss how these findings relate to the aforementioned literature and conclude.

2 Relevant literature

2.1 Goals, round numbers, and reference dependence

An extensive literature documents reference-dependent preferences: people evaluate outcomes not only in absolute terms but also relative to a reference point, coding them as gains or losses (Kahneman & Tversky, 1979; O'Donoghue & Sprenger, 2018; Tversky & Kahneman, 1991, 1992). For instance, an annual bonus of $5,000 feels very different for a recipient who expected nothing than for one who expected a $10,000 bonus: the first codes it as a $5,000 gain, the second as a $5,000 loss. People tend to dislike losses more than they like equivalent gains (loss aversion) and grow less sensitive to outcomes as they move farther away from the reference point (diminishing sensitivity) (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). Reference dependence has been shown to shape behavior across a wide range of economic settings, including professional sports (Bartling, Brandes, & Schunk, 2015; Berger & Pope, 2011; D. Pope & Simonsohn, 2011; D. G. Pope & Schweitzer, 2011), the housing market (Andersen, Badarinza, Liu, Marx, & Ramadorai, 2022; Genesove & Mayer, 2001), financial investing (Odean, 1998; Shefrin & Statman, 1985), labor supply (Camerer, Babcock, Loewenstein, & Thaler, 1997; Crawford & Meng, 2011; Farber, 2005; Fehr & Goette, 2007), tax compliance (Engström, Nordblom, Ohlsson, & Persson, 2015; Rees-Jones, 2018), and education (Fryer, Levitt, List, & Sadoff, 2012).

Identifying the reference point in a given setting is often an empirical challenge. While existing work has focused on the status quo as a reference point (Barberis, 2013; Kahneman & Tversky, 1979; Kőszegi & Rabin, 2006), a growing body of research demonstrates the relevance of non-status-quo reference points, such as expectations (Abeler, Falk, Goette, & Huffman, 2011; Gneezy, Goette,

Sprenger, & Zimmermann, 2017; Strulov-Shlain & Wellsjo, 2025) and past price highs (Heath, Huddart, & Lang, 1999). Round numbers are a particularly natural non-status-quo reference point: they are salient and cognitively accessible, serving as "cognitive reference points" (Rosch, 1975), and motivate effort in a range of settings. For example, high school juniors falling just short of a round number SAT score are significantly more likely to retake the test than those just achieving one, and professional baseball players exert substantially more effort at the end of a season when their batting average sits just below .300 (D. Pope & Simonsohn, 2011). Round numbers also frequently serve as goals (Allen et al., 2017b). For example, a majority of marathon runners disproportionately hold round number finishing time goals—28% of runners selected an "hour goal" finishing time (e.g., 3:00 or 4:00), 49% an hour or half-hour goal, and over 70% selecting a goal time at least at a 15 or 10 minute interval (Markle et al., 2018).

Goals have long been understood as important levers for inducing effort and performance (see Brandstätter & Bernecker, 2022; Fishbach, 2025, for recent reviews). In the literature, goals motivate by directing attention toward goal-relevant activities, mobilizing and sustaining effort, and prompting the discovery of task-relevant strategies (Locke & Latham, 2002). Goal-setting theory, which synthesizes several decades and hundreds of studies of goal setting in workplaces and the laboratory, has demonstrated that specific, challenging goals increase effort and persistence relative to easy goals or vague exhortations to "do your best" (Locke & Latham, 1990, 2002, 2006). For instance, logging truck drivers who had been instructed to simply "do their best" when loading their trucks carried roughly 60% of the legal weight limit; when assigned the specific and challenging target of 94% of the legal limit, loading rose to about 90% within weeks and remained there for years (Latham & Baldes, 1975).

A defining feature of such goals is their specificity: a specific, measurable target divides continuous performance into "success" and "failure", giving individuals an unambiguous standard against which to evaluate outcomes (Heath, Larrick, & Wu, 1999). This builds on the "levels of aspiration" tradition, in which aspired-to performance levels partition outcomes into subjective successes and failures (Atkinson, 1957; Lewin, 1999; Siegel, 1957). This intuition is the building block in Heath, Larrick, and Wu (1999)'s proposition that goals serve as reference points and "inherit" the properties of the prospect theory value function. Because the value function is steep and convex below the

reference point (i.e., for outcomes short of a goal) and less steep and concave above it (i.e., for outcomes past the goal), motivation is higher for individuals just short of a goal than for those just past it, and rises as one approaches a goal while falling after surpassing it. For example, someone with a goal of completing 30 push-ups values three additional push-ups more when they have only completed 27—erasing the shortfall—than when they have already completed 30 and each further push-up is merely a gain. Wu, Heath, and Larrick (2008) formalize this proposition, and Au and Feldman (2020) replicate its central findings. Running provides some of the clearest field evidence for this account: Markle et al. (2018) show that marathon runner satisfaction with finishing times exhibits loss aversion and diminishing sensitivity with respect to (often round number) time goals, closely following the shape of the prospect theory value function.

The proposition that goals are reference points generates a distinctive empirical prediction. Because marginal benefits are high just below a goal and fall sharply above it, performance should "pile up" or "bunch" just beyond the reference point (Heath, Larrick, & Wu, 1999; Kleven, 2016). Allen et al. (2017b) formally show that loss aversion and diminishing sensitivity, as well as a jump in utility at the reference point, produce this pattern, and document bunching in nearly ten million marathon finishing times, with substantial excess mass just below round number finishing times such as 3 and 4 hours; runners adjust both their pacing throughout the race and their effort over the final kilometers to narrowly surpass these thresholds. Similar patterns of bunching around round numbers and goals appear in online chess ratings (Anderson & Green, 2018), road races across a range of distances (Soetevent, 2022), personal tax returns (Engström et al., 2015; Rees-Jones, 2018), and reported corporate earnings (Burgstahler & Dichev, 1997; Degeorge et al., 1999).

2.2 The dynamics of goal pursuit

The "goal gradient" effect, first documented by Hull (1932), describes how motivation changes over the course of goal pursuit, increasing as an agent approaches a goal and diminishing once it is surpassed: consumers accelerate purchases as they approach loyalty-card rewards (Kivetz et al., 2006), with similar patterns in charitable giving and other domains (Cryder et al., 2013; Dai & Zhang, 2019; Liberman & Förster, 2008; Lukas, 2018). Evidence on the far side of the reference point is thinner but consistent: von Rechenberg, Gutt, and Kundisch (2016) show that contributors

to an online community increase their activity as they approach the threshold for a status badge and sharply curtail effort immediately after crossing it. These patterns are consistent with the predictions of the prospect theory value function.

Notably, however, these settings share a common feature: performance accumulates toward a goal whose outcome remains unrealized. The consumer collecting loyalty stamps and the contributor building activity toward a status badge—like the runner mid-race—are still progressing toward a single outcome, and their current position relative to the reference point directly determines the marginal value of further effort. Many performance contexts instead consist of sequential, separate activities, in which an outcome is realized in each episode. For example, the runner achieves a finishing time in race 1 and evaluates it using that day's value function, while the next race is evaluated afresh, under its own (Imas, 2016). To the extent that runners value performances in this way (each as a separate episode) the value function and the goal gradient make less clear predictions for how motivation changes following a near miss or a narrow win. Clearer dynamic predictions emerge, however, if episodes are linked—for example, if agents hold goals defined *by* past performances, such as previous times or personal bests, or if the experience of surpassing a reference point changes the value function itself. We return to these possibilities and how our data speak to them in Section 5. How motivation responds once an outcome relative to a reference point has been realized is therefore an open theoretical and empirical question.

Two prominent psychological accounts make similar predictions about the dynamic effects of surpassing or falling short of a reference point. Self-efficacy theory holds that experiencing success is a powerful source of self-efficacy (i.e., boosting individuals' beliefs in their own capability to execute the actions required for success), which induces effort and persistence (Bandura, 1982). By this account, narrowly surpassing a goal should *increase* subsequent motivation and performance, as success strengthens individuals' perceptions of their own ability to succeed again. The literature on "choking under pressure" documents that performance can deteriorate precisely when incentives and self-imposed pressure are highest, as heightened self-focus and anxiety can distract and disrupt execution (Mesagno & Beckmann, 2017; Yu, 2015). To the extent that narrowly falling short of a reference point concentrates psychological pressure, this account predicts performance decrements for those approaching an as-yet-unachieved threshold. Surpassing it should relieve that pressure,

implying improvement thereafter.

Related work characterizes how individuals regulate motivation across repeated performance episodes. Emotional responses to goal progress govern whether individuals persist, coast, or disengage over the course of pursuit: performing better than expected generates positive affect that can lead individuals to "coast" and reduce subsequent effort (Louro, Pieters, & Zeelenberg, 2007). Perceived progress on a focal goal can likewise license reduced effort or a shift of attention toward alternative goals (Fishbach & Dhar, 2005, 2007), and when pursuing a series of progressive goals, focusing on what remains increases motivation relative to focusing on what has been achieved (Koo & Fishbach, 2010). Fishbach (2025) provides a recent review of this literature.

Recent empirical work examines how narrowly missing or narrowly surpassing a threshold affects subsequent behavior. In academia, early-career researchers who narrowly missed a prestigious NIH grant performed better on several measures of career success than narrow winners despite the substantial loss of resources, though they were also more likely to quit academia (Wang et al., 2019). In competitive sport, online chess players increase effort as they approach both their personal best and round number ratings, and sharply reduce participation and performance after surpassing them (Anderson & Green, 2018). In this online chess environment, however, these changes in effort were only measured over the course of several hours. González-Díaz et al. (2024) extend this analysis to the universe of officially rated classical chess games—an analysis of more than 250,000 players and 22 million games—and show that male and female players respond differently around their personal bests: the performance of female players declined while that of men improved as they neared this reference point. They find both men and women increased participation approaching personal bests but find mixed effects after players surpassed them. Because personal bests update upon being surpassed, they could not track how these effects evolved over longer horizons. In running, Burdina and Hiller (2021) find that marathoners finishing within five minutes above a round number time (i.e., narrowly missing it) are more likely to improve in their next marathon than those finishing within five minutes below it, while Soetevent (2022) finds that performances in the region of round number times have no effect on subsequent participation or performance among Dutch road racers. Both analyses, however, are vulnerable to a mechanical confound: faster times—such as those just below rather than just above a round number—are inherently less likely to be improved upon, both

because of regression to the mean and because faster, more experienced athletes have less room to improve.

2.3 Dynamic effort effects in organizations

Reference points also affect effort and dynamic behavior in firms. Bank managers who lose a leading position in a sales tournament apply more effort to regain it than similar managers who were never leaders, because second place represents a loss for those who have already led (Gutierrez, Obloj, & Frank, 2021). The behavioral theory of the firm holds that firm performance is also evaluated relative to aspiration levels (such as a target return on investment)—targets that function much like reference points, with performance below aspiration triggering search and effort, and performance above it inducing organizational slack (Cyert & March, 1963; March, 1988). Relatedly, non-linear incentive schemes built around discrete targets shape employee effort and firm performance. Firms engage in "earnings management," both moving accounting entries across periods and distorting effort, pricing, and sales strategy to surpass targets and benchmarks. The result is bunching in firm outcomes: reported earnings cluster just above zero, prior-year earnings, and analyst forecasts, and sales cluster just above quotas and targets (Burgstahler & Dichev, 1997; Degeorge et al., 1999; Pierce et al., 2025).

These distortions have consequences across periods: clearing a benchmark in one period has implications for effort and performance in the next. For example, firms that meet benchmarks in one period face incentives to sustain performance strings in subsequent periods (Beneish, 2001; Myers et al., 2007). Salespeople have been shown to distort effort dynamically in response to sales targets, such as quotas and ceilings: they time deals around targets and accounting period ends—"pulling in" sales from future periods when short of a quota and "pushing out" sales when they have already met it—and apply low effort when the target is far away or unreachable (Oyer, 1998). Salespeople also reduce effort after surpassing these targets, and together these dynamics lead to substantial reductions in overall revenue (Misra & Nair, 2011; Pierce et al., 2025). In these settings, however, targets carry material stakes—bonuses, valuations, promotion and firing decisions—so strategic and psychological responses are difficult to separate, and subsequent-period behavior often reflects intertemporal substitution (e.g., deals pulled forward or pushed back) rather than changes

in motivation per se.

2.4 Open questions

Taken together, this work leaves several questions unresolved. Theoretically, the clearest predictions about motivation over the course of goal pursuit come from settings in which effort accumulates toward a single, as-yet-unrealized outcome. When performance instead consists of sequential, separate activities, each evaluated on its own, it is less clear how performance relative to a reference point shapes motivation in subsequent episodes. Empirically, evidence on the far side of a reference point remains thin, and these settings often constrain analyses to short time horizons or even only the next performance, leaving open how these effects evolve over longer time horizons and whether effort is diverted to other pursuits. Evidence from organizational settings, as noted, also cannot cleanly separate material from intrinsic stakes, or motivation from the reallocation of output across periods.

Our data allow us to address several of these open questions. Adolescent athletes compete repeatedly against a fixed, salient reference point that carries no material stakes and does not update once surpassed. This and the scale and structure of our data, described in the following section, allow us to trace the consequences of narrowly missing or narrowly surpassing a reference point across a range of outcomes and time horizons. We then reconcile these results with models of reference-dependent preferences, showing how the standard model can be extended to sequential activities by allowing a reference point to carry less weight once it has been surpassed.

3 Data and Institutional Setting

3.1 High School Track in the United States

We use a novel dataset of track times for high school athletes in the United States. Nearly 1.2 million boys and girls participate in high school track and field in the United States annually, making it the most popular high school sport by combined participation in the country (National Federation of State High School Associations, 2025). Material rewards, such as cash prizes, are prohibited for high school athletes (Florida High School Athletic Association, n.d.; National Federation of State

High School Associations, n.d.), and only 2.1% of performances in our dataset meet the lowest recruiting standard for U.S. college athletics scholarships (the most notable external reward in this setting).[1] Athletes participate in competitions or "meets" over the course of an athletics season, with the indoor season typically running from December to March, and the far more popular outdoor season running from March until June. Each meet provides athletes with an opportunity to compete against other runners in a range of individual events, where individual times and placements for each event are recorded and are the primary outcomes of interest for athletes. Athletes therefore have multiple opportunities within a season to participate, perform, and improve. Importantly, round-number performances, the reference points of interest in our investigation, largely do not coincide with selection criteria for college scholarships, participation in higher-level competition, or other external rewards.[2]

3.2 The dataset

Our data come from Athletic.net, the largest repository of high school athletic outcomes in the United States. As of August 2026, the platform had data on approximately 21.7 million athletes and 221 million individual athlete entries (i.e., outcomes from individual track and field events). For context, 15,311 high schools posted track outcomes on Athletic.net between 2009 and 2019, compared with the 17,052 high schools registered with the National Federation of State High School Associations (the U.S.'s national body that coordinates state high school athletics) for boys' track and field in 2018-19 (National Federation of State High School Associations, 2019). This sample thus represents a large portion of high school athletic outcomes in the United States. Event organizers and coaches upload results from track and field meets onto the platform, producing a standardized set of data, that includes an athlete's ID, race time, finish position, and the date and name of the competition. Athletes on the platform have unique identifiers that link their results across meets. Importantly, the significant majority of race times (93.1%) are recorded using "fully automatic timing" (FAT), i.e., machine-recorded to the 0.01 second.

We focus on times for the boys and girls 800-meter, 1600-meter, and 3200-meter races uploaded

[1] See Next College Student Athlete (n.d.) and (National Collegiate Athletic Association, 2020).

[2] See Next College Student Athlete (n.d.), National Federation of State High School Associations (2021) and Appendix Section A2

onto the platform between 2009 and 2019. This allows us to avoid periods affected by COVID-19-related shutdowns. We scrape these data from the universe of athlete profiles on Athletic.net with at least one performance in any of the above events. These data include 9,116,097 time entries for 716,629 athletes across 1,733,916 distinct athlete-seasons. We focus on these three races as they each, with the exception of the girls 3200 meters, have more than 1 million performance records; have at least one round number time (such as 2 minutes, 3 minutes, etc.) in the interquartile range of finishing times; and provide some scope for athletes to modulate performance in order to exceed round number times (see Appendix Figures A8- A12 and Figure 1 for the distribution of finishing times for each).[3] Our primary analysis focuses on the boys 1600-meter race, and on reference dependence relative to a 5-minute time. Moreover, a "5-minute time" is "round" by running standards. Allen et al. (2017b) show that time thresholds where the left-most digit changes (such as from 4:59 to 5:00) create the most bunching of times, and that multiples of 5 minutes tend to show more reference dependence than other minute thresholds. Finally, 5 minutes lies in a part of the distribution that suggests it should be motivating for a large proportion of runners (see Figure 1 for the distribution of finishing times in this race).[4] We additionally repeat the analysis with the boys and girls 800-meter and 3200-meter races, and the girls 1600-meter race at a set of round numbers in each race.

3.3 Data restrictions and sample

We measure performance over the course of a running season for athletes in the above races. Therefore, it is important to have accurate data for each race that an athlete runs within a season. For our primary analyses, we thus exclude all athlete-seasons with missing times or date information for any race in that season. We also exclude athlete-seasons when an athlete ran a non-metric distance race in the same season that could serve as a substitute event (for example, the one-mile race in addition to the 1600-meter race), athlete-seasons with outlier times, duplicate entries, and "hand times" (see

[3]By contrast, athletes competing in sprint events, such as the 100- or 200-meter races, may not have scope to narrowly surpass round number times, given that requires controlling the final performance to within several hundredths of a second, a very fine margin over which athletes arguably have very limited control. This ability to modulate effort to surpass a round number outcome is a necessary condition to generate bunching in the distribution of finishing times.

[4]40% of all finishing times are within 15 seconds of 5 minutes. However, only 25% of runners run faster than 5 minutes in their first race. This suggests that this reference point is achievable for a large proportion of runners participating in this event, but that it poses a challenge. Heath, Larrick, and Wu (1999) and Locke and Latham (1990) propose that goals affect behavior when they are both achievable and challenging.

Appendix Table A1 for the specific exclusion criteria and the number of observations excluded).[5] This leaves a final dataset of 831,722 times for the boys 1600 meters (41.7% of all recorded times for this race). We use this sample for most analyses in this paper, but, for robustness, we repeat our primary analyses on the full sample and on a set of samples using different exclusion criteria.

Table 1: Descriptive Statistics for Boys 1600 meters Samples

	Final Sample		All Complete Times	
	Mean	Standard Deviation	Mean	Standard Deviation
Year	2015.70	2.77	2015.24	2.94
Month	7.74	1.15	7.79	1.13
Grade	10.43	1.08	10.31	1.19
Indoor Run (Indoor=1, Outdoor=0)	0.10	0.30	0.08	0.27
Number of Races per Athlete-Season	2.38	1.81	3.04	2.30
Days between Races	17.10	19.19	14.46	16.79
Time (Mean) (MM:SS.00)	5:15.99	0:30.48	5:19.44	0:37.20
Time (10th Percentile) (MM:SS.00)	4:40.27	-	4:40.28	-
Time (25th Percentile) (MM:SS.00)	4:53.48	-	4:54.34	-
Time (50th Percentile) (MM:SS.00)	5:11.44	-	5:13.73	-
Time (75th Percentile) (MM:SS.00)	5:34.76	-	5:39.03	-
Time (90th Percentile) (MM:SS.00)	5:58.16	-	6:04.85	-
Seasonal Records (MM:SS.00)	5:16.41	0:30.57	5:19.68	0:35.83
Change in Time (SS.00)	-02.53	11.07	-02.44	22.71
Observations	831,722		2,088,983	

Notes: This table reports the means and standard deviations for observable characteristics of athletes and performances in both our Final Sample (applying all exclusion criteria) and the Full Sample (only removing races where we could not recover reliable performance outcomes). "Change in time" refers to the average change in finishing time between consecutive races for the same athlete (e.g., between race 1 and 2 for an athlete i).

Table 1 provides summary statistics for this final sample and for the full set of finishing times for which we could recover a complete finishing time. Both samples are qualitatively similar on most dimensions. The final sample, however, includes slightly faster times on average, and the number of races per athlete is lower (reflecting the fact that having at least one recording error in a season is more likely when athletes have more entries within a season). For athletes in the final sample, 81.1% of times fall between 4:40 and 6:00, with a mean improvement in time from one race to the next of 2.53 seconds. The vast majority of races take place between March and May, reflecting the popularity of the outdoor season, but there is also substantial participation during December, January and February in the indoor season, with 18.6% of athletes participating in some indoor races (Figure A3). We treat a "season" as beginning in December and ending in August, but also

[5]"Hand times" include times that are recorded by an appointed official using a stop-watch as each athlete crosses the finish line. These are recorded to the 10th of a second and are not as precise as Fully Automatic Times (FATs).

run our analyses excluding any athletes that ran the indoor season for robustness. On average, athletes compete with relatively short intervals between races, with the bulk of consecutive races (83.75%) taking place within 3 weeks of each other, and 22.3% of races occurring on a 7-day cycle (Figure A4).

4 Results

To investigate the effect on future motivation of near misses and narrow wins around round numbers, we must first establish that round numbers serve as reference points in this setting. We do so by testing for "excess mass" in the distribution of finishing times just below round numbers, such as the 5-minute mark in the boys 1600-meter race. We then test for discontinuities in subsequent performance and participation between athletes who narrowly surpass versus narrowly miss these round number times. Following this, we examine how these discontinuities differ depending on the time horizon of subsequent effort and performance and explore several potential mechanisms for our effects. Finally, we consider alternative mechanisms and identification threats.

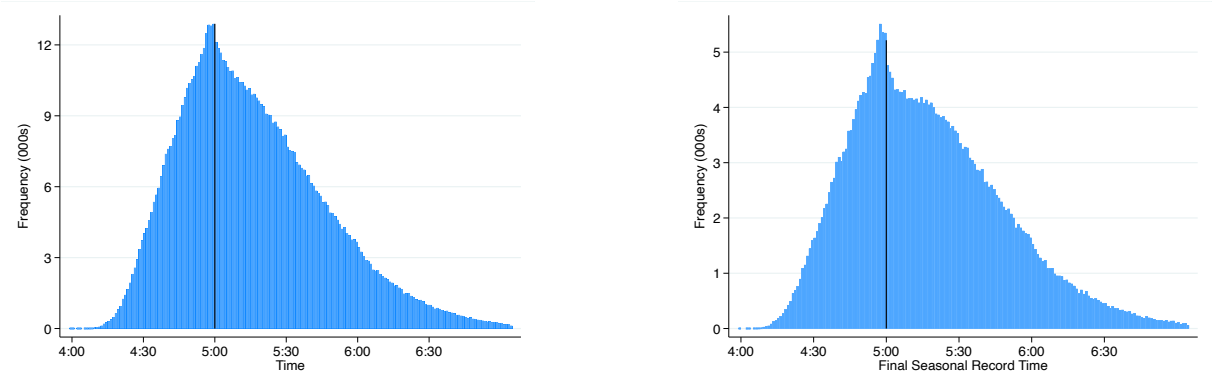
4.1 Bunching in finishing times and seasonal records at round numbers

We first investigate if finishing times are reference-dependent by testing for excess mass below the 5-minute reference point. Figure 1 shows the distribution of all finishing times (left panel) and end-of-season record times (right panel) for the boys 1600-meter race. End-of-season records are defined as the fastest time an athlete achieved in an athletic season. The black line divides the distribution into "successes" and "failures" relative to 5 minutes. We observe noticeable excess mass in both distributions just to the left of this threshold. For instance, for end-of-season records, there are 26,429 instances between 4:55:00 and 4:59.99 compared to 22,584 instances between 5:00:00 and 5:04.99.

To estimate the percentage of excess mass to the left of 5 minutes, we employ a procedure used by Chetty et al. (2011) comparing the actual distribution with a counterfactual distribution of finishing times, i.e., the distribution in the absence of reference dependence at 5 minutes.[6] We

[6]See the Appendix for details on the procedure and robustness exercises. For one use, see Allen et al. (2017b) who employed this procedure to estimate excess mass in the distribution of marathon finishing times.

Figure 1: Distribution of Finishing Times and Seasonal Records in the Boys 1600-meter



(a) Finishing Times
$n = 831,722$

(b) End-of-Season Record Times
$n = 249,693$

Notes: Panels a) and b) show the distribution of finishing times and end-of-season records, respectively, for the boys 1600-meter race. This includes observations in the final sample in 1-second bins.The black line marks 5 minutes and the number of observations in the bin immediately below 5 minutes.

estimate 11.2% excess mass in finishing times ($t = 22.09$, $p < 0.001$) and 17.4% excess mass in end-of-season records ($t = 22.51$, $p < 0.001$) (Table A2). The large amount of excess mass only occurs at 5 minutes and not other time thresholds (see "placebo test": Appendix Figures A6 and A7), and is robust to specification decisions (Appendix Table A4). We find similar but somewhat less pronounced patterns of excess mass in other events, including the girls 1600-meter, and boys and girls 800-meter and 3200-meter races (Appendix Table A3 and Appendix Figures A8-A12). For example, for the girls 1600-meters, we observed 5.1% and 8.2% excess mass in finishing times ($t = 9.00$, $p < 0.001$) and end-of-season records ($t = 7.53$, $p < 0.001$) around 6 minutes (Figure A11). These results highlight the pervasiveness of reference dependence at round number times for high school runners.

4.2 Improvement and persistence near a round number

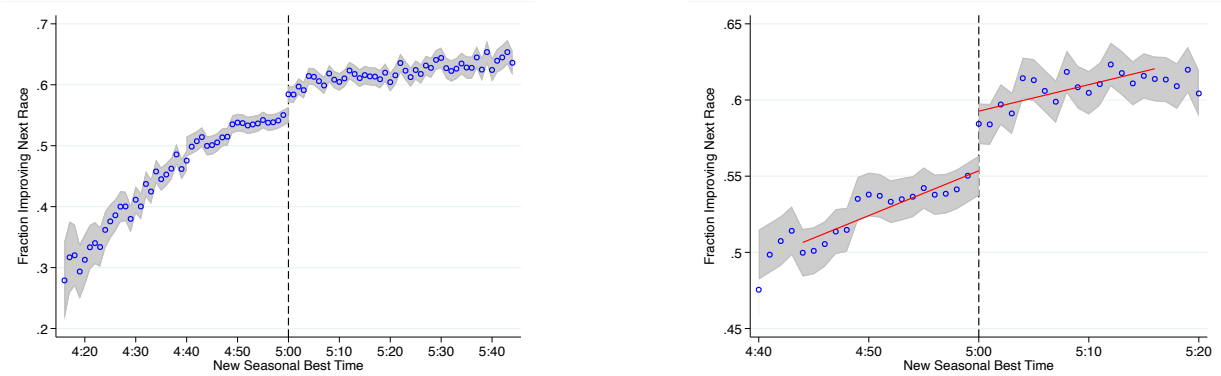
4.2.1 Likelihood of improvement

The observed bunching patterns demonstrate that round numbers, such as 5 minutes in the boys 1600-meter race, serve as reference points for high school athletes. But how does narrowly missing or achieving a reference point affect subsequent performance? To investigate this, we analyze how performance changes in the race following a seasonal record, i.e., an athlete's fastest time that season. The first race in a season is always a seasonal record and subsequent finishing times are

seasonal records only if they surpass the previous fastest time that season. We restrict the analysis to seasonal records in order to control for "seasonal bests" in analysis, as they can serve as alternative reference points to round numbers and are followed by mechanical performance reductions through regression to the mean (Anderson & Green, 2018; González-Díaz et al., 2024). We also focus on these observations because the psychological effects of success and failure are plausibly greater for "best performances", given their relative importance to athletes (Anderson and Green (2018), Markle et al. (2018); see Section 4.4.1 for further discussion). Figure 2 shows the likelihood of improving in the next race (on the y-axis) given a seasonal record time. Each dot represents seasonal bests binned to the second, and the dotted line represents a seasonal record of 5 minutes. Across all seasonal bests, boys improved 58.1% of the time in their next race. The overall upward slope in Figure 2, Panel (a) indicates that improvement becomes more difficult for faster runners. Focusing on the region around 5 minutes, Figure 2 also shows a stark discontinuity in the likelihood of improvement precisely at a seasonal record time of 5 minutes. Boys who narrowly surpassed 5 minutes were notably less likely to improve compared to those who just fell short of 5 minutes. For instance, boys achieving a seasonal best of 5:00.XX improved 58.5% of the time, compared to the 55.0% for boy who ran a best time of 4:59.XX ($t = 3.58$; $p < 0.001; CI : [0.015, 0.053]$).

To estimate the size of this effect, we use a regression discontinuity (RD) approach (Calonico, Cattaneo, & Titiunik, 2014; Imbens & Lemieux, 2008). This involves estimating a linear relationship between the probability of improvement and seasonal best time either side of 5 minutes, and estimating the "drop" in improvement at 5 minutes. The size of the regression "window" (how many seconds either side of 5 minutes are included in the regression) is determined using an optimal bandwidth selection procedure (Calonico et al., 2014; Imbens & Kalyanaraman, 2012). We estimate a 4.0 percentage point decrease in the likelihood of improvement at 5 minutes ($t = 8.89, p < 0.001; CI : [3.08\%, 4.92\%]$). To put this magnitude in context, consider that improvement rates decline steadily as runners get faster: a boy with a seasonal record of 5:32 (the 28th percentile) improves in his next race 63% of the time, compared with 59% for a boy running a 5:00 seasonal record (the 65th percentile). The 4.0 percentage point drop at 5 minutes is thus comparable to the improvement gap associated with a 37 percentile difference in running performance, but arises between athletes separated by a single second.

Figure 2: Probability of Improvement in the Next Race for the Boys 1600-meter



(a) Wide Angle: 4:15 to 5:45

(b) Zoom in: 4:40 to 5:20

Notes: Panels (a) and (b) show the probability improvement in the next race as a function of new seasonal best times for boys in the 1600-meter race. Seasonal best times are binned at the second. Each dot represents the proportion of races in which athletes improved in the next race following a seasonal best time in the corresponding bin, and the grey shaded area represents the 95% confidence interval for this improvement rate. The red line in panel (b) shows the best-fitting line for a piecewise linear regression, estimating a linear function either side of 5 minutes (within an estimated optimal bandwidth) and a discontinuity at 5 minutes.

This sharp reduction in the likelihood of improvement at 5 minutes is highly robust. We observe similarly large effects, often with higher precision, when we relax the sample exclusion criteria to include all finishing times (and a range of alternative exclusion criteria; Appendix Table A8). Our effects are also robust to including race and athlete fixed effects and a set of athlete-race control variables in the regression (Appendix Table A7), and to variations of the RD specification, including a bias correction procedure (Appendix Table A7; Calonico et al. (2014)) and varying the order of the polynomial or the size of the "bandwidth" of the regression window (Appendix Figure A13). We also conduct a "Placebo RD" on all finishing times for which we have sufficient data to estimate the RD (this includes finishing times between 4 minutes and 27 seconds and 6 minutes and 46 seconds) and do not observe a discontinuity at any threshold except 5 minutes (see Figure A14 for details). Finally, our result cannot be explained by potential missing observations from our dataset. Missing observations in an athlete's record *reduce* the estimated drop in performance at 5 minutes and make our analysis conservative (See Appendix Subsection A4.1 for more details).

4.2.2 Other measures and races

We next show similar negative effects of "narrow wins" compared to "near misses" on other margins of performance, using the same regression discontinuity approach described in the previous section.

All of our results show a decrease in performance at the discontinuity, and in what follows, we report the estimated decrease in that measure at the discontinuity (see Table 2 for a summary). We find a 0.36 seconds decrease in improvement in *absolute time* in the next race (i.e., the difference in absolute finishing time between the subsequent race and the present race; $t = 3.51$; $p < 0.001; CI : [0.16, 0.56]$; Appendix Figure A20), a 3.4 percentage point decrease in the likelihood of improvement over the course of the remaining athletic season ($t = 7.82$; $p < 0.001; CI : [2.55\%, 4.25\%]$; Appendix Figure A21), and a 0.36 seconds decrease in improvement in absolute time in seconds over the remaining season ($t = 3.50$; $p < 0.001; CI : [0.16, 0.56]$; Appendix Figure A22)

Boys narrowly surpassing 5 minutes are thus significantly less likely to improve in the next race and over the course of the athletic season. However, does surpassing this threshold also affect their *persistence*, or likelihood of running additional races? We investigate the effect of sub-5-minute times on whether or not boys choose to participate in any additional races that season (as opposed to quitting, and not participating in any further races that season) and on the number of additional races they choose to run. We find a 2.0 percentage point reduction in the likelihood of running another race that same season ($t = 3.93$; $p < 0.001; CI : [1.01\%, 2.99\%]$) and on average 0.056 fewer additional races run ($t = 3.01; p < 0.01; CI : [0.019, 0.093]$; Appendix Figures A23 and A24). Thus, narrowly surpassing a round number time also decreases motivation to participate, and not only motivation as it pertains to performance.

Table 2: Discontinuities in Improvement and Participation: Boys 1600-meter, 5 Minutes

	Effect Size	t-statistic	Average Outcome: 4:59.XX	Average Outcome: 5:00.XX	n (sample size in regression window)
Fraction Improving: Next Race	0.040	8.47	0.550	0.585	161,675
Time Improvement: Next Race	0.360	3.51	1.038	1.287	162,461
Fraction Improving: Rest of Season	0.034	7.82	0.749	0.771	206,765
Time Improvement: to End of Season Best	0.362	3.50	4.898	5.120	165,008
Participation: Any Other Race that Season	0.056	3.01	1.455	1.471	203,516
Participation: Number of Races More	0.020	3.93	0.595	0.600	190,767

Notes: This table reports the details of estimates from the regression discontinuity procedure described in Section 4.2, Regressing outcomes following a seasonal record in the boys 1600-meter race on a dummy variable for times of 5 minutes, and a linear function of times either side of 5 minutes within an optimal bandwidth (Calonico et al., 2014; Imbens & Lemieux, 2008). Each row represents a different outcome and thus a different test. Column 1 presents the discontinuity estimate (positive values indicate negative jumps from surpassing 5 minutes, such as reductions in improvement), column 2 the t-statistic from this regression, and column 5 the size of the sample used in the regression (within the optimal bandwidth). Columns 3 and 4 compare the naive outcome differences for observations in the 1-second bins below and above 5 minutes.

We extend our analysis to other boys events, the 800- and 3200-meter races, and also to a set of girls events, the 800-, 1600-, and 3200-meter races. For these races, we select relevant round

number times as those that are "round" (at minute intervals, or half minute intervals for shorter races), that are relatively "aspirational" (such that they are on the "faster side" of the distribution), and for which we have a minimum number of observations in bins around this reference point (in order to identify any effects). When we investigate the effects of narrowly surpassing round number times on the same margins of performance and participation in these events, we find largely similar results. We find a reduced likelihood of improving in the next race for the boys and girls 3200-meter races, the girls 1600-meter race, and in the boys 800-meter race (Appendix Table A11). These effects are smaller than in the boys 1600-meter race. We do not observe an effect in the girls 800-meter race. However, this race does not contain a threshold time rounded to the minute;[7] that this affects motivation less is consistent with less reference dependence for "less round" times (e.g., Allen, Dechow, Pope, and Wu (2017a)). While the negative effect of narrowly surpassing a round number time on other margins of improvement and persistence replicate less reliably across the other events, we find evidence for all six main effects (on our measures of performance and persistence) in the boys 3200-meter race and five out of six in the girls 1600 meters (Appendix Table A11).

4.2.3 Addressing threats to identification

The regression discontinuity approach relies on the assumption that observations immediately either side of the 5-minute threshold are "as good as randomly assigned." We rely in part on features of the setting to support this assumption. First, it is difficult for athletes in these races to precisely control their finishing times. The margins on either side of 5 minutes in the boys 1600-meter race are in terms of seconds and 10ths of seconds. Strategically allocating effort over several minutes in a dynamic race environment sensitive to other runners' strategies, weather, and other physical conditions make this precise control unlikely. Moreover, unlike in long distance running events (such as 10 kilometers, half-marathons or marathons), runners in the events analyzed in this paper do not typically consult personal watches or other pacing devices, nor do these races have "pace-setters" (athletes tasked with keeping a prescribed pace to help others calibrate their timing). However, if precise control of finishing times *were* possible, given reference dependent preferences we would expect to see very few finishing times slightly slower than 5 minutes, as runners strongly prefer to avoid these times. Instead, we find *more* seasonal record times in the bins 0.1 and 0.2 seconds

[7] The only "round time" in a an aspirational and dense part of the distribution is 2 minutes and 30 seconds.

slower than 5 minutes than in the corresponding bins faster than 5 minutes, and approximately the same number for 0.5 second bins (Figure A18). However, when we test for differences in subsequent performance for athletes within 0.5 seconds of 5 minutes, we observe the same effect of significantly worse performance in the next race for those narrowly surpassing 5 minutes (Figure A17). This shows that while self-selection is unlikely at these small margins, we still observe negative future performance effects.

To further account for possible selection at the 5-minute boundary, we test for discontinuities in other observable characteristics at 5 minutes using the same RD strategy as in our primary analysis Imbens and Lemieux (2008). Ideally, to support the validity of our main effects, we should not find discontinuities in other characteristics of athletes and races at this threshold. Appendix Table A10 and Figure A19 show that characteristics of both athletes and athlete-races—such as athletes' school grade, the month of the race, and performance prior to the focal race—remain unchanged at the threshold.

To further test for robustness to selection, we include different sets of control variables in the regression discontinuity specification. First, to ensure that the effect is not driven by different types of athletes concentrating either side of 5 minutes, we include athlete fixed effects in the RD regression. This identifies the effect of surpassing 5 minutes "within an athlete" (i.e., testing improvement in the next race for an individual athlete who surpassed 5 minutes compared to if that same athlete did not). This should completely account for selection at the athlete level. We find no change to the estimated negative effects of surpassing 5 minutes on improvement and participation with the inclusion of athlete fixed effects (Appendix Table A7).

To account for selection in which *races* are concentrated either side of 5 minutes, we then include race fixed effects in the RD specification. Additionally, we include several other sets of control variables in our specifications. Across all of the robustness checks, the estimates remain largely unchanged and highly significant (Appendix Table A7).

Finally, if athletes were less motivated after surpassing 5 minutes for reasons other than the effect of surpassing this reference point (in a way that cannot be captured by the above observables) then observations just below 5 minutes should display other signatures of lower motivation. For

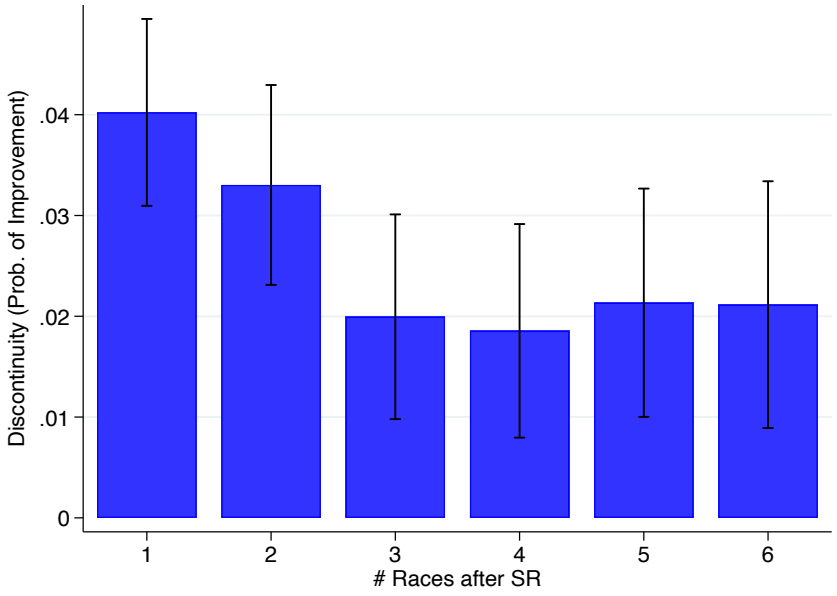
instance, they might perform worse in *other* athletic events. Instead, we find that boys narrowly surpassing 5 minutes do not perform worse in other events that they participate in, such as the 800 meters and 3200 meters, and might even perform *better* (See Section 4.4.2). This suggests that the negative performance and persistence effects of surpassing 5 minutes in the boys 1600-meter are specific to this focal event, and not capturing some spurious form of "general demotivation" in this part of the distribution.

4.3 The longer term effects of round-number performances

4.3.1 Next race vs number of races in the future

We next explore how the effect of a round number performance on motivation changes as an athlete gets further away (chronologically) from the initial performance. First, we compare the discontinuity in improvement in the next race (the main result) to the discontinuity in improvement *two races* into the future (i.e., did the athlete run faster than their initial time in either of the two races following the initial performance), three races into the future, and so forth. This allows us to explore how the effect on future performance evolves over an athlete's season. We find that the discontinuity in improvement at 5 minutes is most pronounced for improvement in the next race, and diminishes approximately monotonically as athletes move further away from the initial performance (Figure 3). Boys in the 1600-meter race are 3.3 percentage points (pp) less likely to improve following a sub-5-minute time two races into the future (compared to 4.0 pp for the next race), and approximately 2 pp less likely to improve up to three races into the future. After three races, the effect appears to stabilize at approximately 2 pp, remaining significant up to six races in the future.

Figure 3: Discontinuity in Improvement by Number of Races in the Future



Notes: Each bar represents the size of the discontinuity in the probability of improvement at 5 minutes in the nth race after a seasonal record (SR) time (for boys achieving a seasonal record in the 1600-meter race). Lines represent 95% confidence intervals of the discontinuity.

4.3.2 No effect on the next season

Next, we test whether the negative effects of surpassing round-number times on motivation persist in the longer term. To do so, we use the same RD technique and test for discontinuities in participation and performance in the *next* athletic season, given an athlete's end-of-season record *this* season.[8] We thus test whether boys who end the season with a Seasonal Record (SR) of 4:59.XX are less likely to participate or improve in the next season compared to boys who end with a SR of 5:00.XX. We find that narrowly surpassing a seasonal record of 5 minutes has no effect on next season's participation: the likelihood of returning the following track season and the number of races an athlete participates in the following track season are not affected (Appendix Table A13 and Figure A30). Boys' seasonal record time next season and their first race time next season are also unchanged at the 5-minute threshold. However, we find a small effect on the probability that an athlete improves next season (compared to the present season): those narrowly surpassing 5 minutes this season are 1.78 percentage points less likely to improve next season ($t = 2.22, p = 0.027, CI : [0.21\%, 3.35\%]$).

[8]We restrict to boys in grades 9-11 as we have data on their participation in the following season. Boys in grade 12 are due to leave the high school track system the following year and thus we cannot capture their participation and performance in the following season.

This effect is robust to multiple specifications, but is relatively small compared to the main effect of within-season performance (4.0 pp), and is less robust.[9] Taken together, these findings suggest that the negative effect of surpassing a round-number time on motivation diminishes in the longer-run, and is modest at best in the following season.

4.4 Exploring the improvement effect

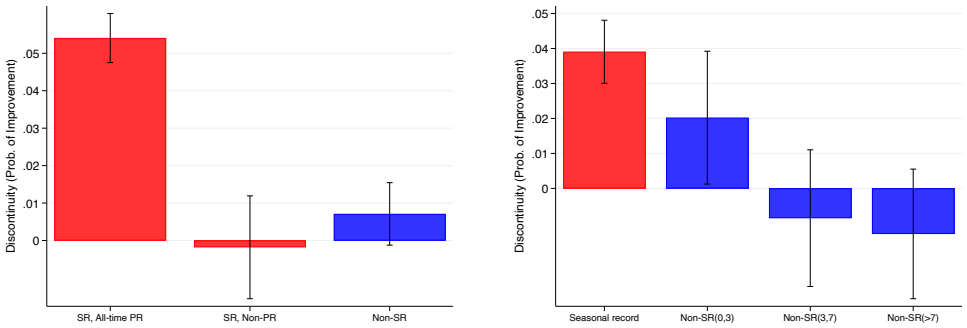
4.4.1 Personal records vs not

To try to understand the mechanism driving reduced improvement after an athlete achieves a round number time, we investigate several features of the effect. First, we test if it matters whether the athlete had previously achieved the round number time or not. Colloquially, the significance of "round number performances" relies on whether or not an athlete has *ever* surpassed a certain round number time. For instance, *Track and Field News*, a historically popular athletics magazine that registers all first-time sub-4-minute mile performances in the United States, only records the *first* time an athlete surpasses 4 minutes. Completing this milestone once is sufficient to join what they call the "Sub-4:00 Miler's Club." We thus investigate whether or not surpassing 5 minutes in the boys 1600-meter has different effects on future motivation whether it is the first time an athlete achieved that time, or whether they had previously surpassed it. To do so, we perform the same regression discontinuity analysis separately for three sets of observations: i) Seasonal Record (SR) times that are *also* all-time personal record times (PRs; i.e., the first time an athlete achieved a given time that season or in any previous season); ii) SR times that are *not* all-time PRs (the first time an athlete achieved a given time *that* season, but had already achieved it in a previous season); iii) Non-SR times (athletes had previously achieved these times that same season). Figure 4 panel (a)) shows the size of the discontinuity in improvement in the next race for these three sets of observations. When athletes had previously achieved a given time, either that season or in a previous season, the discontinuity in improvement at 5 minutes was approximately 0.[10] By contrast, narrowly surpassing an all-time PR of 5 minutes led to a 5.4 percentage point reduction in the probability of improvement ($t = 16.16, CI : [4.8\%, 6.0\%]$). The discontinuity we observe for

[9]The estimate is consistent in magnitude across specifications, and significant at $\alpha = 0.05$ in 3 of 4 specifications, with $p = 0.091$ in the non-significant test. However, it does not survive multiple-hypothesis-testing corrections.

[10]For SRs that were not all-time PRs, $\beta = -0.2\%; t = -0.26, CI : [-1.6\%, 1.2\%]$; for non-SRs, $\beta = 0.7\%; t = 1.67, CI : [-0.1\%, 1.5\%]$

Figure 4: Discontinuities in improvement for Seasonal Records (SRs) compared with Non-SR times.



(a) All-time PRs vs SRs vs Non-SRs (b) SRs vs Gap to SR

Notes: Each bar represents the size of the discontinuity in the probability of improvement at 5 minutes in the next race for boys in the 1600-meter race, with different inclusion criteria for the analysis. In Panel (a), each bar represents improvement discontinuities following different classification of performances, including all-time personal records (PRs), seasonal records (SRs), or performances that are neither (non-SR). Panel (b) shows discontinuities following SRs and Non-SRs of varying the distance in time of the focal performance from the athlete's SR. Lines represent 95% confidence intervals of the discontinuity.

seasonal records is thus fully driven by instances where athletes had *never* previously surpassed 5 minutes. This suggests that surpassing a round number for the first time could have specific motivational effects compared with future round number performances.

Investigating these effects across athletes' careers, we find the effect for all-time PRs is larger in seasons that are later in athletes' high school careers (i.e., in their third and fourth seasons, compared with their first or second; Figure A27). This suggests that the demotivating effect of surpassing 5 minutes might be greater for athletes who have been trying to surpass this threshold for longer. At this stage of an athlete's career, PRs might be more meaningful goals.[11]

Finally, when we decompose non-SR times by how close they are to the seasonal record, we observe a small and slightly significant discontinuity in improvement for finishing times near to the seasonal record (less than 3 seconds from the SR); the discontinuity becomes insignificant (and negatively signed) for times further from the SR (Figure 4, panel b.; Appendix Table A12). This suggests that the psychological effect of the 5-minute reference point on performance might diminish as an athlete's best time gets further from this reference point.

[11] Athletes in their fourth athletic season break their personal records at less than half the rate they did in their second season, for example.

4.4.2 Substitution to other events

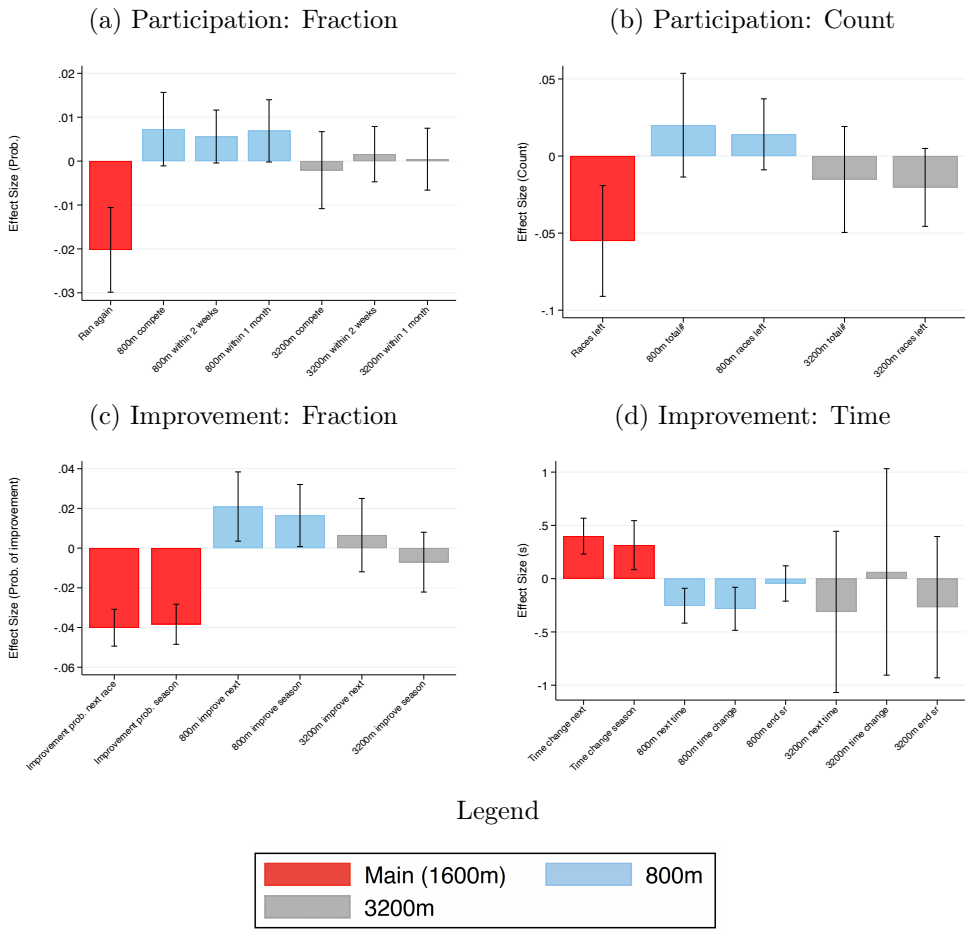
It is possible that athletes achieving a round-number time in a given event respond by shifting their focus to other events (Louro et al., 2007; Shah, Friedman, & Kruglanski, 2002). For instance, having just surpassed 5 minutes in the 1600-meter race for the first time, an athlete might subsequently direct more effort towards other races, such as the 3200-meter race. This potential "substitution of effort" to other events might partly explain the discontinuity in improvement in the focal event, as athletes divert resources towards alternative goals. To examine this possibility, we test the effect of sub-5-minute performances in the boys 1600-meter on participation and performance in other events. We exploit the fact that middle-distance high school athletes typically compete in multiple events: 65% of boys competing in the 1600-meter race also compete in at least one of the 3200-meter or 800-meter races (58.7% participate in the 800-meter and 51% in the 3200-meter). We thus investigate whether performances around the 5-minute reference point in the 1600 meters generate discontinuities in participation and performance in these other events.

Figure 5 shows the magnitude of the discontinuity for several measures of participation and performance in these events (blue bars for the 800-meter and grey bars for the 3200-meter), following a sub-5-minute SR in the boys 1600-meter. These are benchmarked against effects in the 1600-meter itself (red bars). Overall, we do not find evidence of systematic changes in participation or performance in these alternative events. Participation rates in both events and performance in the 3200-meter race are unaffected by sub-5-minute times in the 1600 meters. There is, however, a small but significant effect on performance in the boys 800 meters: those narrowly achieving 5 minutes in the 1600-meter race are 0.25 seconds faster ($t = 3.05, p < 0.01, CI : [0.09, 0.42]$) in their next 800-meter race and are 2 pp more likely to improve relative to their previous 800-meter performance ($t = 2.35, p = 0.017, CI : [0.9\%, 3.3\%]$) (Table A14). These effects are smaller than our main results for improvement in the 1600-meter itself and slightly less robust.[12]

These results suggest that athletes may adjust their effort on the intensive margin in the 800-meter race, possibly applying more effort in that event after achieving a sub-5-minute time in the 1600-meter. Could this substitution of effort explain the main results, such that athletes shift focus

[12]The result for finishing times in the next 800-meter race remains significant and the magnitude stable across specifications, but results for the probability of improvement lose significance in some specifications.

Figure 5: Shifting Focus: Discontinuities in improvement and participation for other events, 800 meters and 3200 meters, compared with main effects



Notes: Each bar represents the size of the discontinuity estimated from the regression discontinuity procedure described in Section 4.2 for different outcomes. All discontinuities are with respect to outcomes following seasonal record performances in the boys 1600-meter race, and with respect to discontinuities at 5 minutes. Panel a) estimates discontinuities for the fraction of athletes participating, panel b) for the count of races, panel c) for the fraction of athletes improving, and panel d) for the time improvement in seconds. Red bars reflect discontinuities estimated for outcomes in the boys 1600-meter race, light blue for the boys 800-meter, and gray for the boys 3200-meter.

to the 800-meter and consequently perform worse in the 1600 meters? To test this, we take two approaches. First, we run the original RD on improvement in the 1600-meter race while controlling for participation and performance in other events (the boys 800 and 3200-meter races). Second, we separately estimate the effect for athletes who do and do not participate in other events. In both cases, the main effects remain statistically and qualitatively equivalent (all point estimates for improvement exceed 3.5 pp and are significant at $p < 0.01$), suggesting that "shifting focus" to other events cannot account for the main discontinuities in participation and performance in the boys 1600-meter.

4.4.3 Risk taking

The results thus far show that narrowly achieving a sub-5-minute time has a negative effect on performance in the following race *in the aggregate*. However, it is possible that this aggregate result is driven by changes in "risk taking" at this threshold. Athletes who narrowly surpass 5 minutes might take more risks in subsequent races, leading to improvements for athletes for which the "gamble" paid off, and to decrements for those for which it did not. This increased risk-taking would thus lead to greater dispersion in outcomes as athletes surpass 5 minutes. We thus test for discontinuities in the *variability* of performance at 5 minutes for the boys 1600-meter race. We find no evidence of a discontinuity in the standard deviation of finishing times at seasonal records of 5 minutes (Appendix Figure A33), suggesting no clear change in risk-taking at this threshold.

Additionally, we conduct a quantile regression of time improvements (in seconds) on the 5-minute threshold, using the same RD strategy for different deciles of time improvement. This allows us to separately test for discontinuities in improvement for athletes that improved by a large margin (higher deciles), those that did not improve, and those that performed worse in the next race (lower deciles). We find meaningful and significant reductions in improvement after sub-5-minute seasonal record times in 7 out of 9 deciles of time improvement (Appendix Table A15 and Appendix Figure A34). This suggests that the negative effects on time improvement are relevant to a large proportion of the outcome distribution, and not captured solely by "bad gambles" (i.e., not solely driven by sharp reductions in the lower deciles). While the negative performance effects of surpassing 5 minutes are lower for deciles where athletes improved the most (which we would expect with changes in risk-taking), the overall pattern of results provides limited evidence that risk-taking drives the reduction in average improvement.[13]

4.4.4 Exhaustion

Could the negative effect on performance be driven by physical exhaustion from the effort exerted to narrowly surpass five minutes? First, we do not observe a discontinuity in time improvement *leading*

[13]There is no discontinuity in dispersion of time improvement at 5 minutes, which we would expect with changes in risk-taking. Moreover, we observe the negative effect on performance in the 40th, 50th, and 60th percentiles of improvement, which would not be the case if the effect were fully driven by changes in risk-taking. Relatedly, while wee see null and negative effects for deciles where athletes improve the most, the change in improvement is not monotonic as we move from the lowest to highest improving deciles.

up to the focal event: athletes who narrowly surpassed five minutes do not appear to have applied discontinuously greater effort to do so. Second, if exhaustion were driving the effect, it should fade with recovery time: athletes racing again within days of a sub-5:00 performance should be more affected than those racing weeks later. We therefore estimate the discontinuity in improvement separately by the interval between the focal race and the next one. The effect is approximately the same whether the next race is within one week, between one and two weeks, or between two and three weeks, and only slightly smaller—though still significant—beyond three weeks (Appendix Figure A35).[14] Physical exhaustion is therefore unlikely to explain the effect.

4.5 A model of dynamic reference dependence

We have shown that runners bunch at five minutes, that those who narrowly surpass it improve less in their next race, and that both effects are stronger for personal and seasonal records than for ordinary finishing times. Reference dependent preferences predict the first of these, bunching in finishing times at round numbers (Allen et al., 2017b). Whether they can account for the other two is less clear. We therefore simulate a population of runners using a model of reference dependent preferences over three separate races to test whether it can reproduce these results. We also test a modification to this model, in which the weight on reference dependence reduces after the reference point has been surpassed—capturing the idea that a threshold matters less once a runner has cleared it.

The value function is given in Equation (1), which determines the value runner i places on a time t in period p. Here t is defined as the number of seconds faster than the slowest possible time, so that faster times take a *higher* value of t. There is a kink in the value function at $t = r$: it is steeper for times that fall short of the reference point r than for times that surpass it. This is due to loss aversion, represented by λ, and generates *bunching* in times as runners apply additional effort to avoid losses. The parameter $\eta_{ip} \geq 0$ is the weight the runner places on reference-dependent utility.

[14] Only 3.4% of next races occur within two days of the focal race, the window in which incomplete recovery, and therefore direct exhaustion effects, would be most severe. This subset is too small to drive the overall effect, and in any case the point estimate there is actually *smaller* than the main effect, though is not statistically different from it ($\beta = 2.4\%, t = 0.95, p = 0.343, CI : [-2.6\%, 7.5\%]$)

$$V_{ip}(t) = \begin{cases} t - \eta_{ip}\lambda(r-t), & t \leq r \\ t + \eta_{ip}(t-r), & t > r \end{cases} \quad \Longrightarrow \quad V'_{ip}(t) = \begin{cases} 1 + \eta_{ip}\lambda, & t \leq r \\ 1 + \eta_{ip}, & t > r \end{cases} \tag{1}$$

The two specifications differ only in how η_{ip} evolves between periods. In the *baseline* model, $\eta_{ip} = \eta$ in every period. In the *deflated* model, a runner who has surpassed the reference point places reduced weight on it thereafter, such that $\eta_{ip} = \eta^{PR}$ where $\eta^{PR} \leq \eta$:

$$\eta_{ip} = \begin{cases} \eta^{PR}, & t_{iq} > r \ \text{ for some } q < p, \\ \eta, & \text{otherwise,} \end{cases}$$

The deflated model nests the baseline as the special case $\eta^{PR} = \eta$.

Following Allen et al. (2017b), a runner's realized time t_{ip} is the point at which marginal cost equals marginal benefit, $V'_{ip}(t)$. Marginal cost is a power function,

$$C'_{ip}(t) = A_{ip}t^{\beta}, \tag{2}$$

where β governs the steepness of the cost curve and A_{ip} is an inverse ability parameter: stronger runners have lower A_{ip} and so face lower costs at any t. Between races, A_{ip} changes such that runners become faster on average but with substantial dispersion, partially calibrated to the race-to-race changes in the empirical data (see the Appendix for further details).

Figure 6 shows how the two models diverge. It follows two runners between period 1 and period 2: one who falls two seconds short of five minutes in period 1 (5:02, a "near-miss"; Panel A) and one who surpasses it by two seconds (4:58, a "narrow-gain"; Panel B). Each runner's period-1 time is set where their marginal cost curve c'_1 meets the horizontal marginal value function, marked by the black square. In period 2, A_{ip} changes and the marginal cost curve (hereafter, "cost curve") shifts. The change in A_{ip} is normally distributed, and the dotted curves c'_{2a}–c'_{2g} illustrate the range of cost curves a runner might face in period 2, evenly spaced across one standard deviation either side of the mean change. That mean change is c'_{2d}, slightly to the right of c'_1 since runners improve on

average. Period-2 time is then set where the realized cost curve crosses the marginal value function, as in period 1.

In Panel A the near-miss runner faces the same marginal value function in period 2 as in period 1 (the black horizontal line), having never surpassed r. In Panel B the narrow-gain runner faces either that same function, under the baseline model, or the lower one implied by η^{PR} (in teal). Under the baseline model, then, whether a runner improves depends entirely on the shift in marginal cost, and because that shift is distributed identically for both runners, so is their probability of improvement. Integrating over the full distribution of cost changes, both improve with probability 56.6%—as does a runner at any other period-1 time, on either side of the threshold, at any values of λ, η, and β.[15] The baseline model thus predicts no discontinuity in improvement probability at 5 minutes, and no parameterization of it can produce one.[16]

The deflated model, however, breaks the equivalence. Facing the lower marginal benefit function, the narrow-gain runner responds to the mean change in marginal cost (given by c'_{2d}) by running 4:58.03—three hundredths of a second slower than in period 1, despite improved ability—where the same change moves the baseline runner from 4:58 to 4:57, a 1-second improvement. Integrating over the full distribution of potential marginal costs, his improvement probability falls to 49.8%, 6.8 percentage points below the near-miss runner's. This shows how this model can produce a discontinuity in the improvement probability. The figure also shows why the baseline model cannot reproduce the negative discontinuity in time improvement. In fact, it produces a discontinuity of the wrong sign, predicting that narrow-gain runners improve by *more* than near-miss runners, where our data show the reverse. The appendix describes the mechanism behind this pattern. By contrast, the deflated model produces the negative discontinuity in time improvement.

The intuition here is relatively simple. If five minutes matters as much to a runner after he has achieved it as it did before, he remains motivated to clear it again, and will do so in any race in which his ability has not declined. Only if the threshold loses importance once surpassed will

[15] Under the baseline model the map from ability to realized time is non-decreasing and identical in every period, so a runner improves if and only if his ability improves. Improvement probability is therefore $\Phi(\mu/\sigma)$ for every runner, where μ and σ are the mean and standard deviation of the period-to-period ability change.

[16] The exception is the small group of runners whose period-1 time is exactly at r. A modest ability improvement leaves them at r again and so with a time change of exactly zero, making it ambiguous whether to count them as improving.

runners slack off after a narrow win.

This intuition bears out in the full simulation results. The baseline model generates bunching at 5:00 but produces no discontinuity in improvement probability and a time-improvement discontinuity of the wrong sign. By contrast, the deflated model reproduces the bunching along with a negative discontinuity in improvement probability, and, at fitted parameters, a negative discontinuity in time improvement of approximately the same magnitude as the empirical data. We fit both models by simulated method of moments, targeting six moments: bunching in finishing times and in seasonal records, and the discontinuities in improvement probability and in time improvement following each. The deflated model fits substantially better and returns a deflation factor of $\eta^{PR}/\eta = 0.347$, well below one; because the deflated model nests the baseline at $\eta^{PR}/\eta = 1$, this is evidence against the baseline model. The baseline model fit, by contrast, drives reference dependence towards zero, returning $\eta = 0.010$ and $\lambda = 1.080$: almost no weight on the reference point and almost no loss aversion. In switching reference dependence off, the fitted baseline model loses the bunching it was introduced to explain, a result inconsistent both with previous work and with our data. Table A17 reports these results, and the appendix reports sensitivity to λ, η, and the deflation ratio. These results are also robust to alternative value functions: one in which the runner's target time is realized with noise, and one with diminishing sensitivity around the reference point, as in related work (Allen et al., 2017b).

Deflation in the model is triggered by a runner's *first* crossing of the reference point. This points to a broader extension of the standard model of reference dependence: across sequential, separate activities, what a reference point does to effort depends on whether it has already been surpassed.

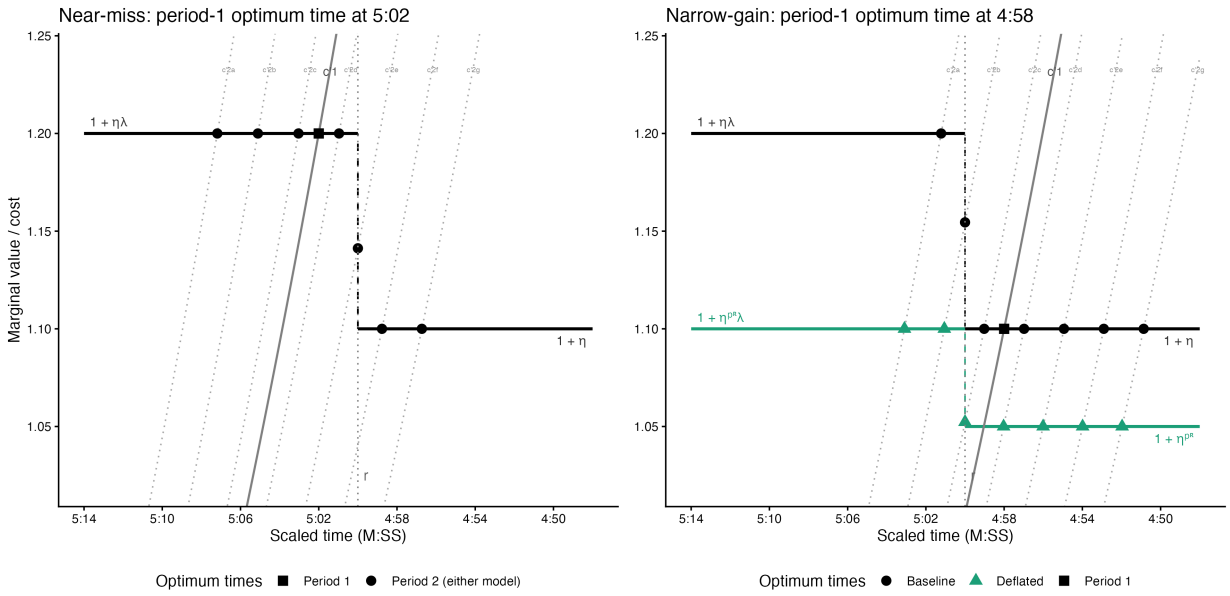


Figure 6: Illustrative optimum response to a family of period 2 ability changes. Panel A: a runner whose period 1 optimum sits 2 seconds short of r (5:02, "near-miss"). Panel B: a runner 2 seconds past it (4:58, "narrow-gain"). Grey dotted lines are period 2 cost curves c'_{2a}–c'_{2g}, one per ability change $\delta \in \{-5, -3, -1, 1, 3, 5, 7\}$ seconds, centered on the mean change $\delta = 1$ and spanning roughly $\pm$ one standard deviation; the solid grey line is the first-race cost curve, c'_1. Black horizontal steps are the undeflated marginal benefit curve—the only curve relevant in Panel A, since r is never surpassed there, and the baseline curve in Panel B. The teal step in Panel B is the deflated marginal benefit curve, which applies only once r has been surpassed. Marginal benefit levels are labeled directly on each curve. Points mark the resulting period 2 optima for each change, colored and shaped by model as in the legend; where baseline and deflated coincide, only one marker style is shown.

5 Discussion & conclusion

We find that athletes who narrowly surpass a round-number time are both less likely to improve performance and less likely to persist than those who narrowly miss one. Boys achieving seasonal record times just under 5 minutes in the 1600 meters were 4.0 percentage points less likely to improve in their next race than boys who just missed it, and improved 0.36 seconds less. These boys were also 3.4 percentage points less likely to improve at any point over the remainder of the season, 2.0 percentage points less likely to race again that season, and ran 0.056 fewer additional races. The pattern replicates in the girls 1600 meters and in the boys and girls 800- and 3200-meter races, and is concentrated among athletes surpassing the threshold for the first time. The effect fades over time, however, and is undetectable a season later. Surpassing a goal thus reduces subsequent striving: what we call "slacking after success."

This effect is not explained by physical exhaustion, regression to the mean, or changes in risk-taking. Research on goals suggests that shifting effort to other pursuits could explain the effect: attaining a focal goal (like a 5-minute time in the 1600-meter race) frees up resources to pursue other goals (like other distance races), shifting effort away from the focal pursuit (Förster, Liberman, & Higgins, 2005; Kruglanski et al., 2002; Shah et al., 2002). However, we find limited evidence of effort substitution to alternative races following a 5-minute seasonal record.

We find that the negative effect on improvement depends on whether 5 minutes was surpassed for the first time or had previously been achieved: the effect is large for all-time personal records and indistinguishable from zero otherwise. This suggests that the motivational significance of a round-number goal diminishes once it has been surpassed. Consistent with this, Appendix Table A6 shows less *bunching* in finishing times below 5 minutes for athletes who had previously surpassed it than for those who had not.

This intuition motivates a modification to the standard model of reference-dependent preferences, in which the weight placed on the reference point attenuates once a runner has surpassed it for the first time. The standard model of reference dependence gives clear predictions for how effort changes *within* an episode whose outcome remains unrealized (Heath, Larrick, & Wu, 1999), but is less clear about how effort should change across sequential, separate episodes, each realized

on its own. Simulations show that the standard model cannot reproduce the discontinuities in improvement we observe. However, allowing the weight on the reference point to fall once it has been surpassed reproduces these discontinuities and the bunching patterns in finishing times and seasonal records (O'Donoghue & Sprenger, 2018; Tversky & Kahneman, 1992).

We thus build on previous theoretical and empirical work by showing how motivation responds to missing or achieving a reference point across sequential, realized episodes, and by extending models of reference dependent preferences to accommodate these effects. This builds on research showing how the reference point itself adapts over time in response to past outcomes (DellaVigna, Lindner, Reizer, & Schmieder, 2017; Thakral & Tô, 2021); we instead hold the reference point fixed and capture adaptation by changing its motivational weight. The present work is also relevant to research that shows that realized and unrealized outcomes carry different implications under prospect theory: unrealized losses generate more risk-taking than realized ones (Imas, 2016). Here, we instead show how realization changes a reference point's importance, applied to effort rather than risk.

Our setting shares several features with organizational contexts. Goals of this kind are common in organizations—sales quotas, revenue milestones, service-level targets. Workers pursue these targets across sequential periods while juggling multiple goals and responsibilities at once: a quota for one product alongside a quota for another, a performance review to prepare, a protégé to train. Our measures of effort, persistence, and substitution likewise map onto outcomes organizational scholars care about, such as productivity, retention, and effort allocation. Together, these parallels raise the possibility that similar dynamics govern goal pursuit inside organizations. Existing work shows that targets do shape behavior dynamically in organizations—output shifts across periods to clear a threshold, and drops off once a target is exceeded—but such data rarely allow effort to be measured cleanly, or psychological responses to be separated from strategic responses to material stakes. Our setting isolates the negative *psychological* effect of surpassing a goal directly, in data of a kind rarely available inside a firm. This suggests that a similar cost operates in organizations, and may underlie some of the patterns in output already documented there.

The present research thus builds on other work highlighting the hidden costs of setting targets to induce effort. Targets by their nature focus effort narrowly, and have been shown to divert

effort from other activities that a firm values (Pierce et al., 2025), reduce net output (Misra & Nair, 2011), and produce a range of further unintended costs, from distorted risk preferences and inhibited learning to unethical behavior and reduced intrinsic motivation (Ordóñez, Schweitzer, Galinsky, & Bazerman, 2009). Slacking after success adds a cost of a different kind to this list: one incurred not when a target is missed or gamed, but when it is achieved.

Five minutes in the boys 1600 meters is a particularly motivating goal: it is highly round, salient, applicable to a large share of runners, and reasonably ambitious relative to the performance distribution (Allen et al., 2017b; Locke & Latham, 1990, 2002, 2006). Reference points that score lower on these criteria might produce smaller declines in motivation once surpassed, which would explain why other events in our data—the girls 1600 meters and the boys and girls 800 and 3200 meters—show smaller and less reliable effects. We do not, however, formally test importance as a moderator. Future work could measure goal importance directly (Hollenbeck & Williams, 1987) and test its impact on the effects we document.

Our setting also offers few alternative goals to 5 minutes. The presence of alternatives—particularly ones that could replace the focal goal, such as a new, faster round number or a personal record—might mitigate the negative effects on motivation that we observe. Here, personal records appear considerably less motivating than 5 minutes: both bunching and the effect on next-race improvement are smaller at personal-record times (Appendix Section A6). And the next comparable round number, 4:30, is both less round and far harder to reach—34.1% of boys surpass 5 minutes in a season, compared with 3.4% who reach 4:30—diminishing its capacity to drive effort once 5 minutes has been surpassed. For instance, only 0.03% of boys who surpass 5 minutes in a season go on to break 4:30 that same season. Other settings might look different. In golf, for example, a player who breaks par (typically 72) can set their sights on breaking 70, a close and still-round successor. Where a successor goal sits close at hand, or renews on a shorter cycle, the demotivating effect we document here may be correspondingly smaller as motivation is renewed by this new target.

We also note several limitations of the present work. First, the discontinuity at 5 minutes establishes that boys who surpass the goal subsequently exert less effort and persist less than boys who narrowly miss it, but it cannot, on its own, tell us which side of the threshold drives the gap: reduced motivation after surpassing the goal, elevated motivation among those still short of

it, or both. Visual inspection and statistical extrapolation of the relationship between performance and our outcomes suggest a decline after surpassing the reference point, but a definitive answer requires a counterfactual for effort in the absence of a reference point altogether. The ideal test would experimentally vary the existence of the reference point itself. Unfortunately, in our data we cannot cleanly derive the counterfactual. Second, we cannot measure substitution to non-athletic activities such as schoolwork, so we cannot rule out that some of the reduced effort we observe is redirected outside the domains we can see. Finally, we cannot cleanly compare the effect across genders: the round numbers we investigate sit at different points in the male and female performance distributions, making comparisons fraught. For example, 5 minutes is only plausible for a small fraction of girls in the 1600 meters.

Surpassing a goal today can reduce effort and persistence tomorrow. Roger Bannister chased a four minute mile throughout his athletic career, and left competitive running four months after achieving it. Our data cannot settle what a goal is worth across the whole arc of its pursuit, but we document that targets that focus effort toward their pursuit can also stall progress once that pursuit concludes.

References

Abeler, J., Falk, A., Goette, L., & Huffman, D. (2011). Reference points and effort provision. *American Economic Review*, *101*(2), 470–492.

Akerlof, G. A., & Yellen, J. L. (1986). *Efficiency wage models of the labor market*. Cambridge University Press.

Allen, E. J., Dechow, P. M., Pope, D. G., & Wu, G. (2017a). *Online appendix: "reference-dependent preferences: Evidence from marathon runners"*. (Online appendix to Allen, Dechow, Pope and Wu (2017), Management Science 63(6))

Allen, E. J., Dechow, P. M., Pope, D. G., & Wu, G. (2017b, June). Reference-dependent preferences: Evidence from marathon runners. *Management Science*, *63*(6), 1657–1672. doi: 10.1287/mnsc.2015.2417

Andersen, S., Badarinza, C., Liu, L., Marx, J., & Ramadorai, T. (2022). Reference dependence in the housing market. *American Economic Review*, *112*(10), 3398–3440.

Anderson, A., & Green, E. A. (2018, February). Personal bests as reference points. *Proceedings of the National Academy of Sciences*, *115*(8), 1772–1776. doi: 10.1073/pnas.1706530115

Ariely, D., Gneezy, U., Loewenstein, G., & Mazar, N. (2009, April). Large stakes and big mistakes. *The Review of Economic Studies*, *76*(2), 451–469. doi: 10.1111/j.1467-937X.2009.00534.x

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review*, *64*(6), 359–372.

Au, N., & Feldman, G. (2020). *Revisiting "goals as reference points": Replication and extensions of Heath et al. (1999)*. OSF preprint. (https://doi.org/10.17605/OSF.IO/WMQTB)

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, *37*(2), 122–147.

Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, *27*(1), 173–196.

Bartling, B., Brandes, L., & Schunk, D. (2015). Expectations as reference points: Field evidence from professional soccer. *Management Science*, *61*(11), 2646–2661.

Beneish, M. D. (2001). Earnings management: A perspective. *Managerial Finance*, *27*(12), 3–17.

Berger, J., & Pope, D. (2011). Can losing lead to winning? *Management Science*, *57*(5), 817–827.

Brandstätter, V., & Bernecker, K. (2022). Persistence and disengagement in personal goal pursuit. *Annual Review of Psychology*, *73*, 271–299.

Burdina, M., & Hiller, S. (2021, October). When falling just short is a good thing: The effect of past performance on improvement. *Journal of Sports Economics*, *22*(7), 777–798. doi: 10.1177/15270025211018247

Burgstahler, D., & Dichev, I. (1997). Earnings management to avoid earnings decreases and losses. *Journal of Accounting and Economics*, *24*(1), 99–126.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, *82*(6), 2295–2326.

Camerer, C., Babcock, L., Loewenstein, G., & Thaler, R. (1997). Labor supply of New York City cabdrivers: One day at a time. *The Quarterly Journal of Economics*, *112*(2), 407–441.

Chetty, R., Friedman, J. N., Olsen, T., & Pistaferri, L. (2011, May). Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from Danish tax records. *The Quarterly Journal of Economics*, *126*(2), 749–804. doi: 10.1093/qje/qjr013

Crawford, V. P., & Meng, J. (2011). New York City cab drivers' labor supply revisited: Reference-dependent preferences with rational-expectations targets for hours and income. *American Economic Review*, *101*(5), 1912–1932.

Cryder, C. E., Loewenstein, G., & Seltman, H. (2013). Goal gradient in helping behavior. *Journal of Experimental Social Psychology*, *49*(6), 1078–1083.

Cyert, R. M., & March, J. G. (1963). *A behavioral theory of the firm.* Englewood Cliffs, NJ: Prentice-Hall.

Dai, H., & Zhang, D. J. (2019). Prosocial goal pursuit in crowdfunding: Evidence from Kickstarter. *Journal of Marketing Research*, *56*(3), 498–517.

Degeorge, F., Patel, J., & Zeckhauser, R. (1999). Earnings management to exceed thresholds. *The Journal of Business*, *72*(1), 1–33.

DellaVigna, S., Lindner, A., Reizer, B., & Schmieder, J. F. (2017, November). Reference-dependent job search: Evidence from Hungary. *The Quarterly Journal of Economics*, *132*(4), 1969–2018. doi: 10.1093/qje/qjx015

Engström, P., Nordblom, K., Ohlsson, H., & Persson, A. (2015). Tax compliance and loss aversion. *American Economic Journal: Economic Policy*, *7*(4), 132–164.

Farber, H. S. (2005). Is tomorrow another day? the labor supply of New York City cabdrivers. *Journal of Political Economy*, *113*(1), 46–82.

Fehr, E., & Goette, L. (2007). Do workers work more if wages are high? evidence from a randomized field experiment. *American Economic Review*, *97*(1), 298–317.

Fishbach, A. (2025). *Goals and motivation*. Situational Press.

Fishbach, A., & Dhar, R. (2005). Goals as excuses or guides: The liberating effect of perceived goal progress on choice. *Journal of Consumer Research*, *32*(3), 370–377.

Fishbach, A., & Dhar, R. (2007). Dynamics of goal-based choice. In *Handbook of consumer psychology* (pp. 611–637). New York, NY: Psychology Press.

Florida High School Athletic Association. (n.d.). *Bylaw 9.9: Amateurism*. Retrieved 2026-09-03, from https://www.fldoe.org/core/fileparse.php/20758/urlt/5.2.pdf (FHSAA Handbook)

Förster, J., Liberman, N., & Higgins, E. T. (2005). Accessibility from active and fulfilled goals. *Journal of Experimental Social Psychology*, *41*(3), 220–239.

Fryer, R. G., Levitt, S. D., List, J., & Sadoff, S. (2012). *Enhancing the efficacy of teacher incentives through loss aversion: A field experiment* (Tech. Rep.). National Bureau of Economic Research.

Genesove, D., & Mayer, C. (2001). Loss aversion and seller behavior: Evidence from the housing market. *The Quarterly Journal of Economics*, *116*(4), 1233–1260.

Gneezy, U., Goette, L., Sprenger, C., & Zimmermann, F. (2017). The limits of expectations-based reference dependence. *Journal of the European Economic Association*, *15*(4), 861–876.

González-Díaz, J., Palacios-Huerta, I., & Abuín, J. M. (2024). Personal bests and gender. *The Review of Economics and Statistics*, *106*(2), 409–422.

Gutierrez, C., Obloj, T., & Frank, D. H. (2021). Better to have led and lost than never to have led at all? lost leadership and effort provision in dynamic tournaments. *Strategic Management Journal*, *42*(4), 774–801.

Hansen, D. G. (1997). Worker performance and group incentives: A case study. *ILR Review*, *51*(1), 37–49.

Heath, C., Huddart, S., & Lang, M. (1999). Psychological factors and stock option exercise. *The Quarterly Journal of Economics*, *114*(2), 601–627.

Heath, C., Larrick, R. P., & Wu, G. (1999, February). Goals as reference points. *Cognitive Psychology*, *38*(1), 79–109. doi: 10.1006/cogp.1998.0708

Hollenbeck, J. R., & Williams, C. R. (1987). Goal importance, self-focus, and the goal-setting process. *Journal of Applied Psychology*, *72*(2), 204–211.

Hull, C. L. (1932). The goal-gradient hypothesis and maze learning. *Psychological Review*, *39*(1), 25–43.

Hutchinson, A. (2018). *Endure: Mind, body, and the curiously elastic limits of human performance.* New York, NY: HarperCollins.

Imas, A. (2016). The realization effect: Risk-taking after realized versus paper losses. *American Economic Review*, *106*(8), 2086–2109.

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, *79*(3), 933–959.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292.

Kivetz, R., Urminsky, O., & Zheng, Y. (2006). The goal-gradient hypothesis resurrected: Purchase acceleration, illusionary goal progress, and customer retention. *Journal of Marketing Research*, *43*(1), 39–58.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, *8*(1), 435–464.

Koo, M., & Fishbach, A. (2010). Climbing the goal ladder: How upcoming actions increase level of aspiration. *Journal of Personality and Social Psychology*, *99*(1), 1–13.

Kőszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, *121*(4), 1133–1165.

Kruglanski, A. W., Shah, J. Y., Fishbach, A., Friedman, R., Chun, W. Y., & Sleeth-Keppler, D. (2002). A theory of goal systems. In *Advances in experimental social psychology* (Vol. 34, pp. 331–376). Academic Press.

Latham, G. P., & Baldes, J. J. (1975). The "practical significance" of Locke's theory of goal setting. *Journal of Applied Psychology*, *60*(1), 122–124.

Lewin, K. (1999). Level of aspiration. *Psychological Review*.

Liberman, N., & Förster, J. (2008). Expectancy, value and psychological distance: A new look at goal gradients. *Social Cognition*, *26*(5), 515–533.

Locke, E. A., & Latham, G. P. (1990). Work motivation and satisfaction: Light at the end of the tunnel. *Psychological Science*, *1*(4), 240–246.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705–717.

Locke, E. A., & Latham, G. P. (2006). New directions in goal-setting theory. *Current Directions in Psychological Science*, *15*(5), 265–268.

Louro, M. J., Pieters, R., & Zeelenberg, M. (2007). Dynamics of multiple-goal pursuit. *Journal of Personality and Social Psychology*, *93*(2), 174–193.

Lukas, M. (2018). The goal gradient effect and repayments in consumer credit. *Economics Letters*, *171*, 208–210.

March, J. G. (1988). The behavioral theory of the firm. In *Decisions and organizations* (chap. 2). Oxford: Basil Blackwell.

Markle, A., Wu, G., White, R., & Sackett, A. (2018). Goals as reference points in marathon running: A novel test of reference dependence. *Journal of Risk and Uncertainty*, *56*(1), 19–50.

Mesagno, C., & Beckmann, J. (2017, August). Choking under pressure: Theoretical models and interventions. *Current Opinion in Psychology*, *16*, 170–175. doi: 10.1016/j.copsyc.2017.05.015

Misra, S., & Nair, H. S. (2011). A structural model of sales-force compensation dynamics: Estimation and field implementation. *Quantitative Marketing and Economics*, *9*(3), 211–257.

Myers, J. N., Myers, L. A., & Skinner, D. J. (2007). Earnings momentum and earnings management. *Journal of Accounting, Auditing & Finance*, *22*(2), 249–284.

National Collegiate Athletic Association. (2020). *Estimated probability of competing in college athletics.* Retrieved 2026-09-03, from https://www.ncaa.org/sports/2015/3/2/estimated-probability-of-competing-in-college-athletics.aspx (NCAA Research. High school figures from the 2018–19 NFHS High School Athletics Participation Survey)

National Federation of State High School Associations. (n.d.). *Sanctioning: Permissible awards for invited schools.* Retrieved 2026-09-03, from https://tools.nfhs.org/Sanctioning/documents/Sanctioning%20-%20Permissible%20Awards%20for%20Invited%20Schools

.pdf

National Federation of State High School Associations. (2019). *2018–19 high school athletics participation survey* (Tech. Rep.). Indianapolis, IN: National Federation of State High School Associations. Retrieved 2026-09-03, from https://www.nfhs.org/media/1020406/2018-19-participation-survey.pdf

National Federation of State High School Associations. (2021, July). *NIL rulings do not change for high school student-athletes.* Retrieved 2026-09-03, from https://nfhs.org/stories/nil-rulings-do-not-change-for-high-school-student-athletes (The NFHS Voice)

National Federation of State High School Associations. (2025). *2024–25 high school athletics participation survey* (Tech. Rep.). Indianapolis, IN: National Federation of State High School Associations. Retrieved 2026-09-03, from https://nfhs.org/resources/sports/high-school-participation-survey-archive

Next College Student Athlete. (n.d.). *Men's college track and field recruiting standards.* Retrieved 2026-09-03, from https://www.ncsasports.org/mens-track-and-field/scholarship-standards

Odean, T. (1998). Are investors reluctant to realize their losses? *The Journal of Finance*, *53*(5), 1775–1798.

O'Donoghue, T., & Sprenger, C. (2018). Reference-dependent preferences. In *Handbook of behavioral economics: Applications and foundations 1* (Vol. 1, pp. 1–77). Elsevier. doi: 10.1016/bs.hesbe.2018.07.003

Ordóñez, L. D., Schweitzer, M. E., Galinsky, A. D., & Bazerman, M. H. (2009). Goals gone wild: The systematic side effects of overprescribing goal setting. *Academy of Management Perspectives*, *23*(1), 6–16.

Oyer, P. (1998). Fiscal year ends and nonlinear incentive contracts: The effect on business seasonality. *The Quarterly Journal of Economics*, *113*(1), 149–185.

Pierce, L., Rees-Jones, A., & Blank, C. (2025). The negative consequences of loss-framed performance incentives. *American Economic Journal: Economic Policy*, *17*(1), 506–539.

Pope, D., & Simonsohn, U. (2011, January). Round numbers as goals: Evidence from baseball, SAT takers, and the lab. *Psychological Science*, *22*(1), 71–79. doi: 10.1177/0956797610391098

Pope, D. G., & Schweitzer, M. E. (2011). Is Tiger Woods loss averse? persistent bias in the face of

experience, competition, and high stakes. *American Economic Review*, *101*(1), 129–157.

Rees-Jones, A. (2018). Quantifying loss-averse tax manipulation. *The Review of Economic Studies*, *85*(2), 1251–1278.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, *7*(4), 532–547.

Shah, J. Y., Friedman, R., & Kruglanski, A. W. (2002). Forgetting all else: On the antecedents and consequences of goal shielding. *Journal of Personality and Social Psychology*, *83*(6), 1261–1280.

Shefrin, H., & Statman, M. (1985). The disposition to sell winners too early and ride losers too long: Theory and evidence. *The Journal of Finance*, *40*(3), 777–790.

Siegel, S. (1957). Level of aspiration and decision making. *Psychological Review*, *64*(4), 253–262.

Soetevent, A. R. (2022, March). Short run reference points and long run performance. (No) evidence from running data. *Journal of Economic Psychology*, *89*, 102471. doi: 10.1016/j.joep.2021.102471

Strulov-Shlain, A., & Wellsjo, A. S. (2025). Goals, expectations, and performance. *Working paper*.

Thakral, N., & Tô, L. T. (2021, August). Daily labor supply and adaptive reference points. *American Economic Review*, *111*(8), 2417–2443. doi: 10.1257/aer.20170768

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

von Rechenberg, T., Gutt, D., & Kundisch, D. (2016). Goals as reference points: Empirical evidence from a virtual reward system. *Decision Analysis*, *13*(2), 153–171.

Wang, Y., Jones, B. F., & Wang, D. (2019, December). Early-career setback and future career impact. *Nature Communications*, *10*(1), 4331. doi: 10.1038/s41467-019-12189-3

Wu, G., Heath, C., & Larrick, R. P. (2008). *A prospect theory model of goal behavior.* (Working paper, University of Chicago)

Yu, R. (2015, February). Choking under pressure: The neuropsychological mechanisms of incentive-induced performance decrements. *Frontiers in Behavioral Neuroscience*, *9*. doi: 10.3389/fnbeh.2015.00019

Appendix

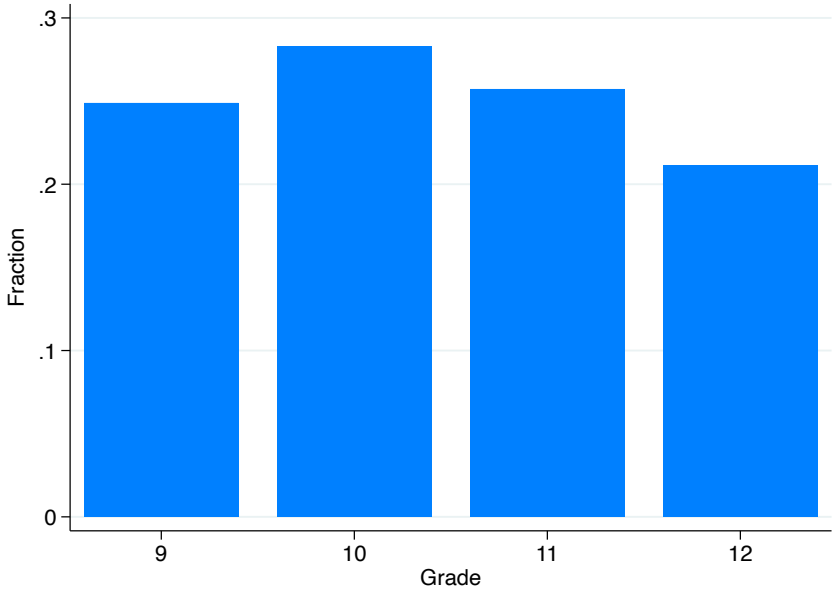
A1 Descriptives

Table A1: Observations excluded in each event for all criteria

	Boys			Girls		
	800m	1600m	3200m	800m	1600m	3200m
Total observations before dropping						
Total times	1,646,352	2,394,008	1,308,364	1,241,046	1,673,072	853,255
Total athlete-seasons	551,953	769,097	462,350	392,664	522,201	283,958
Total athletes	300,374	415,181	265,955	201,326	274,861	152,058
Dropped Times						
Date Missing or Outside 2009-2019	53,494	72,612	44,559	69,796	83,081	48,153
Grade Record Missing or Error	31,543	39,663	17,248	22,175	24,209	7,365
Not in Highschool	65,834	77,726	20,148	57,439	66,804	19,831
Duplicate Entry	62,141	74,129	42,947	47,234	54,278	30,640
Ran the Mile/2 Mile/880y	3,180	254,065	45,585	2,258	146,065	23,152
Time Record Missing or Error	299,999	369,912	253,849	222,669	271,489	166,150
Hand Recorded Time	509,867	648,898	346,918	368,896	440,666	226,046
Outlier Time	17,209	25,524	14,349	15,829	18,293	9,326
Dropped Athlete-Seasons						
Date Missing or Outside 2009-2019	26,245	34,251	22,742	28,064	32,475	20,588
Grade Record Missing or Error	13,721	23,162	10,365	8,768	14,059	3,749
Not in Highschool	26,390	30,286	8,971	22,359	25,522	8,271
Duplicate Entry	12,327	14,523	9,340	9,070	10,542	6,446
Ran the Mile/2 Mile/880y	661	60,339	11,566	406	29,962	4,960
Time Record Missing or Error	75,736	87,243	68,794	53,719	62,203	44,646
Hand Recorded Time	129,609	160,385	93,181	86,754	105,513	57,139
Outlier Time	5,328	9,262	6,446	5,062	7,024	3,996
Totals observations after dropping						
Total times	603,085	831,479	522,761	434,750	568,187	322,592
Total athlete-seasons	261,936	349,646	230,945	178,462	234,901	134,163
Total athletes	181,746	248,506	163,908	120,514	167,219	91,427

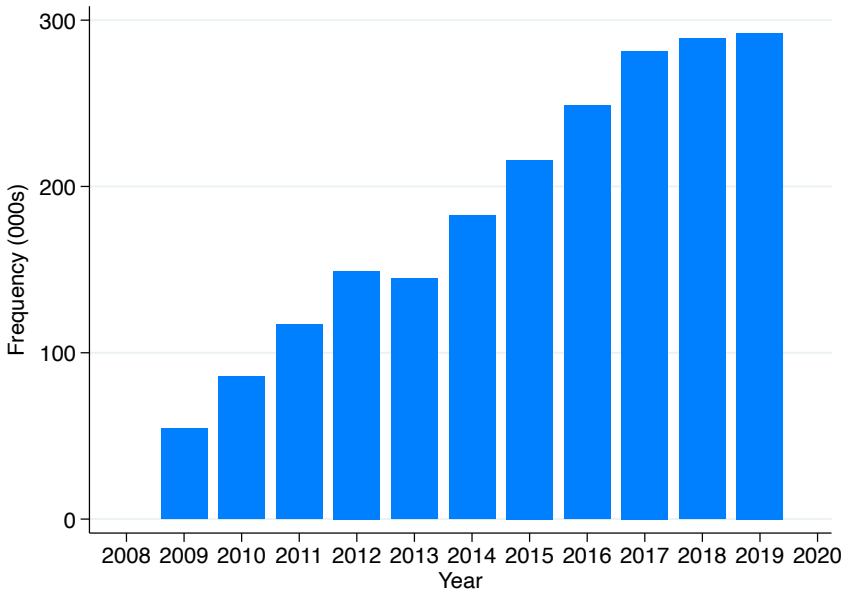
Notes: Times and athlete-seasons excluded under each criterion, by event and gender, before (top panel) and after (bottom panel) applying all exclusion criteria. Categories are not mutually exclusive: an observation flagged for multiple reasons is counted under each, so category counts need not sum to the total excluded.

Figure A1: Distribution of grades of athletes: boys 1600-meter race



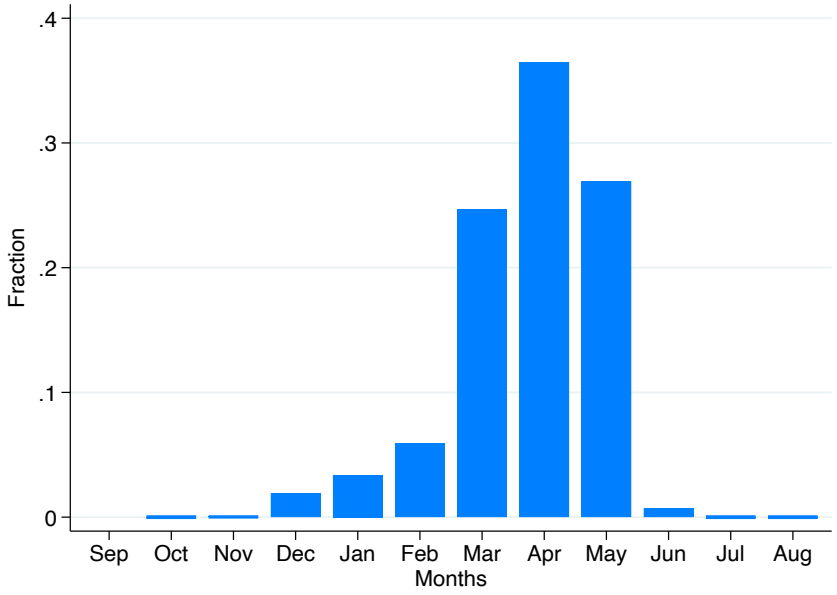
Notes: Distribution of grade level (9th–12th) across all times in the final sample.

Figure A2: Distribution of entries across years: boys 1600-meter race



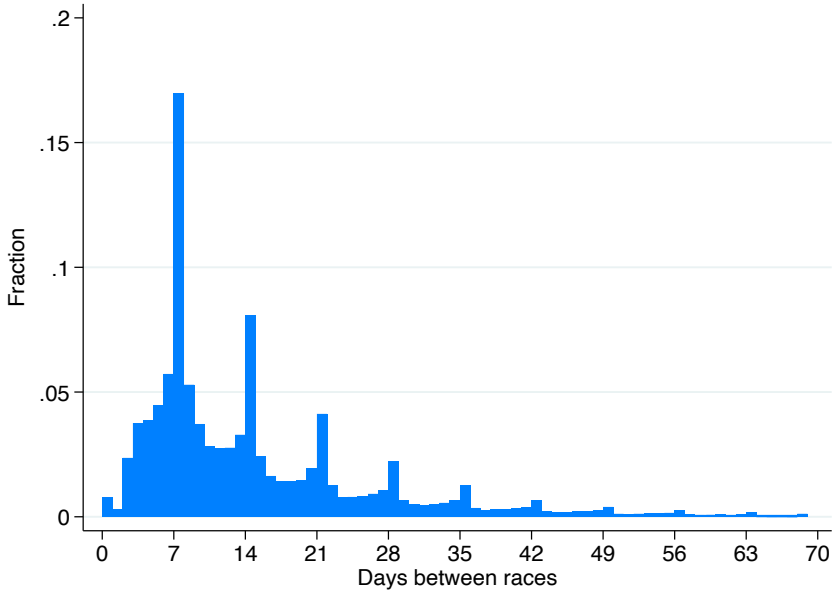
Notes: Distribution of race year across all times in the final sample, 2009 onward.

Figure A3: Distribution of race month: boys 1600-meter race



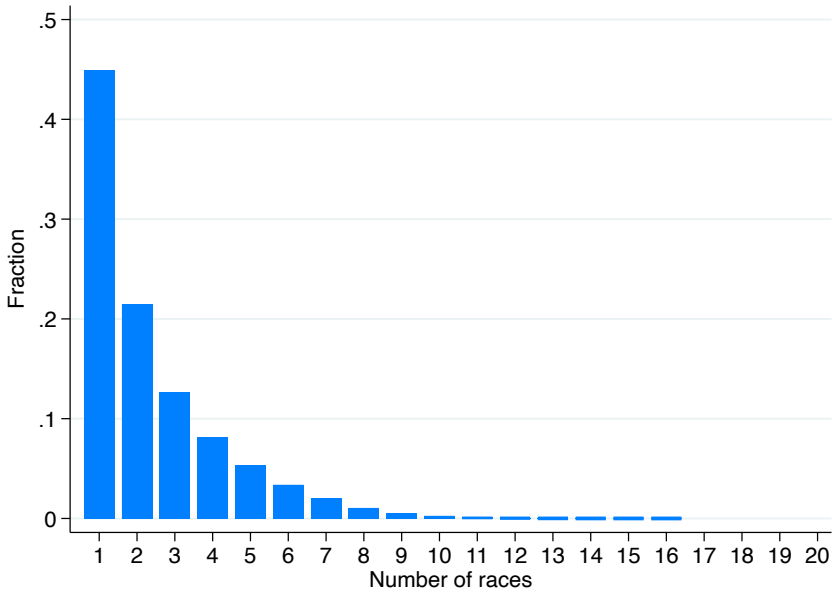
Notes: Distribution of race month across all times in the final sample.

Figure A4: Distribution of days between races: boys 1600-meter race



Notes: Distribution of the number of days between consecutive races for the same athlete in the final sample.

Figure A5: Distribution of number of races per athlete per season: boys 1600-meter race



Notes: Distribution of the number of races run per athlete-season in the final sample, one observation per athlete-season.

A2 State qualifying times

We investigated all state qualifying times published and accessible online for the boys and girls 800-meter, 1600-meter, and 3200-meter races. We elected to store the table online rather than include it in this document as it contains 438 rows: it is available here. Notably, for the events we investigate, only 6 out of 438 published state championship qualifying standards overlapped with round numbers in our data. None of these were 5 minutes in the boys 1600-meter race.

A3 Bunching in finishing times and seasonal records

A3.1 Procedure for estimating excess mass

To estimate the size of excess mass in finishing times and test its statistical significance we require a counterfactual distribution of finishing times over the same bunching region — the distribution we would expect in the absence of bunching induced by the reference points we investigate. To construct this counterfactual and to test for the statistical significance of the excess mass, we apply the procedure developed in Chetty et al. (2011), who used it to test for bunching in the income of Danish taxpayers around kinks in the tax code. Allen et al. (2017b) applied the same procedure to test for bunching in marathon finishing times at round numbers, and our implementation follows their application closely.

Before estimation, we discretize finishing times into quarter-second bins, rounding each time to the nearest 0.25 seconds (so that, for example, a recorded time of 5:00.00 falls in the bin spanning 4:59.875–5:00.125).

To estimate the counterfactual distribution, we fit a quartic polynomial to the density of finishing times in a local window around the reference point, excluding the bunching region itself. In the boys 1600-meter race, the window extends 25 seconds on either side of the reference point, and the bunching region is the 5 seconds immediately below the reference point. As in Chetty et al. (2011), the size of the window and bunching region are chosen by visual inspection of where excess mass appears in the distribution; we additionally test the robustness of our estimates to alternative window and bunching-region sizes (described below). Concretely, for a reference point at 5 minutes, we fit the polynomial to the density between 4:35 and 5:25, excluding the interval between 4:55 and

5:00, the bunching region.

Fitting the polynomial to this local density, excluding the bunching region, and extrapolating it into the bunching region yields a first-pass counterfactual that understates the true counterfactual density: because excess mass in the bunching region is likely drawn from finishers who would otherwise have finished just outside it, excluding that mass from the fit pulls the fitted curve down. Following Chetty et al. (2011) and Allen et al. (2017b), we correct for this by shifting the entire fitted curve upward by a constant so that the total area under the counterfactual, over the local window, equals the total area under the actual density over that same window — avoiding double-counting runners who bunch at the reference point.

The excess mass estimate is the difference between the actual number of finishers in the bunching region and the number predicted by the adjusted counterfactual, expressed as a percentage of the counterfactual. We estimate the standard error of this statistic via a bootstrap procedure with 1000 replications, resampling and re-estimating the full counterfactual-fitting procedure at each replication.

As a placebo test, we repeat this entire procedure — fitting a local counterfactual and estimating excess mass — at every one-second threshold between 4 minutes 23 seconds and 6 minutes 30 seconds, restricting to thresholds with at least 100 observations in the relevant one-second bin and with sufficient data on both sides to support the full 25-second window. This lets us evaluate whether the excess mass we detect at 5 minutes is distinctive relative to the pattern of estimates the same procedure produces at arbitrary, non-reference thresholds.

We apply the same procedure to round-number thresholds in the girls 1600-meter race and the boys and girls 800-meter and 3200-meter races (Appendix Table A3 and Appendix Figures A8–A12). We also test the sensitivity of our excess mass estimates for the boys 1600-meter race to the specification of the bunching procedure: the size of the local window, the size of the bunching region, and the polynomial order used to estimate the counterfactual (Appendix Tables A4 and A5).

Table A2: Excess mass in the boys 1600-meter race at 5 minutes

	Final Sample			
	Actual finishers	Counterfactual finishers	% excess finishers	t-statistic
Finishing Time	65,887	59,259	11.2	22.09
Seasonal Record Time	48,155	42,716	12.7	24.61
End of Season Record Time	27,642	23,543	17.4	22.51
	All Complete Times			
	Actual finishers	Counterfactual finishers	% excess finishers	t-statistic
Finishing Time	153,035	138,739	10.3	34.68
Seasonal Record Time	101,653	90,855	11.9	33.82
End of Season Record Time	51,237	43,079	18.9	33.16

Notes: Actual and counterfactual finisher counts, excess mass (as a % of the counterfactual), and t-statistics from the procedure described in Appendix A3.1, for finishing times, seasonal records, and end-of-season records, in the full and final samples.

Table A3: Excess mass in all events at different thresholds

		Finishing Times				Seasonal Records		End of Season Records	
Event	Discontinuity (mm:ss)	Actual finishers	Counterfactual finishers	% excess finishers	t-statistic	% excess finishers	t-statistic	% excess finishers	t-statistic
Boys 1600 Meter	05:00	65,887	59,259	11.2	22.53	12.7	22.26	17.4	20.61
	10:00	20,333	18,607	9.3	10.73	11.9	10.98	13.5	10.56
Boys 3200 Meter	11:00	30,459	28,813	5.7	8.69	6.5	7.68	9.6	9.88
	12:00	17,459	17,063	2.3	2.96	2.7	2.50	2.7	2.01
Boys 800 Meter	02:00	21,537	20,889	3.1	3.86	3.5	4.23	8.2	6.63
Girls 1600 Meter	06:00	33,980	32,322	5.1	9.00	5.8	8.21	8.2	7.53
	11:00	2,511	2,382	5.4	2.05	2.5	0.92	5.5	1.46
Girls 3200 Meter	12:00	12,294	11,598	6.0	5.28	6.1	4.96	8.2	5.13
	13:00	14,519	14,251	1.9	1.73	1.4	1.16	1.7	1.12
Girls 800 Meter	02:30	25,828	25,416	1.6	2.25	2.2	2.77	3.3	2.80

Notes: Excess mass (%) and t-statistics from the procedure described in Appendix A3.1, applied separately to each event and gender at its own round-number threshold (boys: 800 meters at 2:00, 1600 meters at 5:00, 3200 meters at 10:00; girls: 800 meters at 2:30, 1600 meters at 6:00, 3200 meters at 11:00), for finishing times, seasonal records, and end-of-season records, in the full and final samples.

Table A4: Excess mass in the boys 1600-meter race: different specifications in the Chetty et al. (2011) procedure

Bunching Area	Actual finishers	Counterfactual finishers	% excess finishers	t-statistic
Local Window=10				
1	16,052	15,302	4.9	5.70
2	28,884	27,481	5.1	8.06
3	41,611	39,488	5.4	8.63
4	54,065	51,243	5.5	10.35
5	65,887	63,052	4.5	8.40
Local Window=20				
2	28,884	26,288	9.9	15.88
4	54,065	49,016	10.3	23.02
6	77,405	71,356	8.5	19.34
8	99,575	93,301	6.7	15.91
10	120,687	114,960	5.0	11.81
Local Window=30				
3	41,611	36,531	13.9	21.32
6	77,405	68,907	12.3	27.51
9	110,239	100,069	10.2	30.49
12	141,235	129,308	9.2	27.71
15	169,105	157,447	7.4	22.05
Local Window=40				
4	54,065	45,694	18.3	35.89
8	99,575	86,097	15.7	39.69
12	141,235	123,318	14.5	41.76
16	177,816	157,409	13.0	39.27
20	209,207	188,756	10.8	32.83

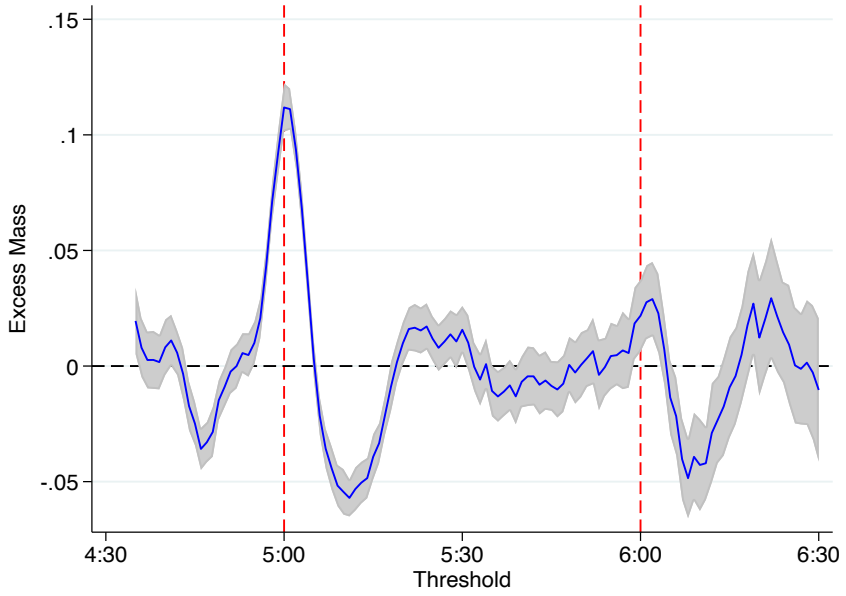
Notes: Excess mass (%) and t-statistics in finishing times at 5 minutes, re-estimated across a grid of local window sizes (10–30 seconds either side of the threshold) and bunching-region sizes (1–6 seconds), holding the polynomial order fixed at the main specification's quartic. Standard errors use a 100-replication bootstrap (a coarser version of the 1000-replication bootstrap used for the main estimates in Table A2, applied here across many specifications for tractability).

Table A5: Excess mass in the boys 1600-meter race: different polynomial orders in the Chetty et al. (2011) procedure

Polynomial Order	Actual finishers	Counterfactual finishers	% excess finishers	t-statistic
3	65,887	59,259	11.2	25.45
4	65,887	59,332	11.0	22.06
5	65,887	60,388	9.1	19.22
6	65,887	60,331	9.2	19.54
7	65,887	60,280	9.3	18.81
8	65,887	60,236	9.4	18.83

Notes: Excess mass (%) and t-statistics in finishing times at 5 minutes, re-estimated for polynomial orders 3 through 8, holding the local window (25 seconds) and bunching region (5 seconds) fixed at the main specification's values. Standard errors use a 100-replication bootstrap, as in Table A4.

Figure A6: Excess mass "placebo test": excess mass at all threshold times, boys 1600-meter race



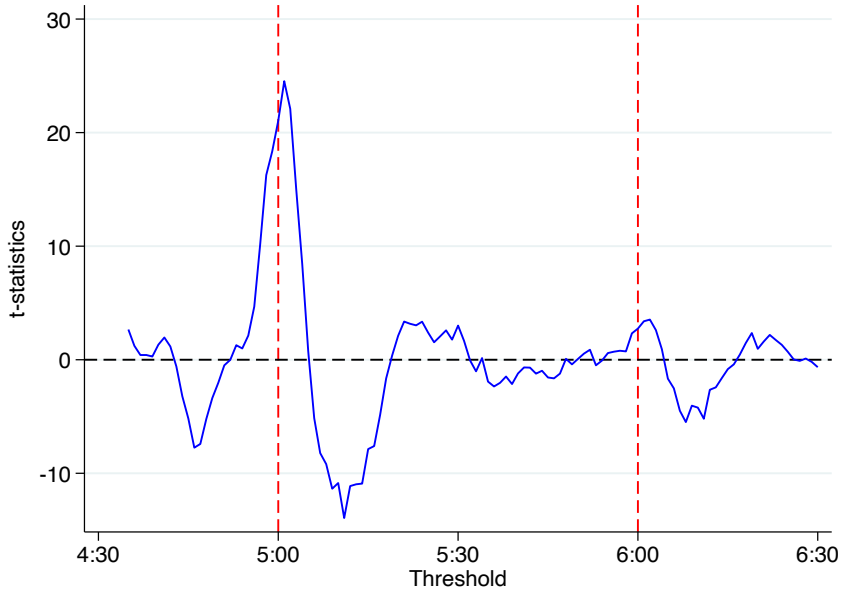
Notes: The results of a "placebo test", estimating excess mass using the Chetty (2011) procedure at all time thresholds (at the level of the second) between 4 minutes and 35 seconds and 6 minutes and 30 seconds — this includes all thresholds for which we have greater than 100 observations for the 1-second bin at the threshold, and at least 20 observations per 1-second bin in the support for the window required to estimate the counterfactual distribution (25 seconds either side of the threshold).

Table A6: Excess mass in the boys 1600-meter race, by previous sub-5-minute performances

Inclusion criteria	Proportion with previous sub-5	% excess finishers	t-statistic	n
No Previous Sub-5 (Same season)	0%	13.2%	20.70	639,439
Previous Sub-5 (Same season)	100%	5.0%	6.46	192,283
1 Previous Sub-5 (Same season)	100%	17.0%	14.65	82,638
2 Previous Sub-5s (Same season)	100%	-2.5%	-1.42	45,514
>=3 Previous Sub-5s (Same season)	100%	-14.6%	-9.33	64,131
5:00<=Prev PR (Season)<5:10	0%	33.6%	29.29	66,195
4:50<=Prev PR (Season)<5:00	100%	17.8%	18.58	73,001
No sub-5 in previous season	0%	8.6%	10.56	596,399
Sub-5 in previous season	100%	4.4%	4.67	114,604
1 Sub-5 in previous season	100%	8.7%	4.04	32,341
2 Sub-5s in previous season	100%	5.1%	1.98	19,878
>=3 Sub-5 in previous season	100%	1.2%	0.84	62,385
5:00<=Prev PR (All time)<5:10	0%	14.7%	7.81	38,139
4:50<=Prev PR (All time)<5:00	100%	12.2%	8.43	46,261

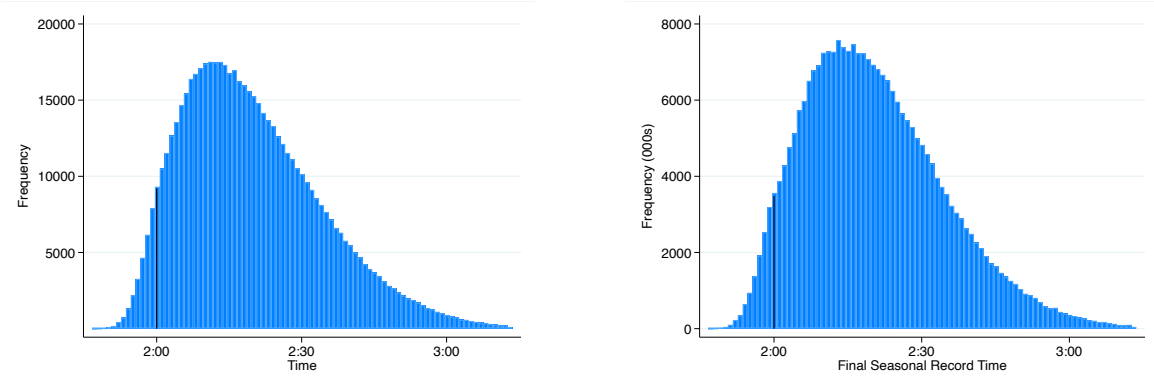
Notes: Excess mass (%) and *t*-statistics in finishing times at 5 minutes, estimated separately for subsamples defined by an athlete's history of sub-5-minute performances: number of prior sub-5 times that season, whether their seasonal record entering the race was just below or just above 5 minutes, and the analogous categories using an athlete's full high-school career ("all-time") rather than season-to-date history.

Figure A7: Excess mass "placebo test": t-statistic for excess mass at all threshold times, boys 1600-meter race



Notes: t-statistics from the same placebo procedure as Figure A6, plotted by threshold.

Figure A8: Boys 800-meter race: distribution of finishing times and seasonal records

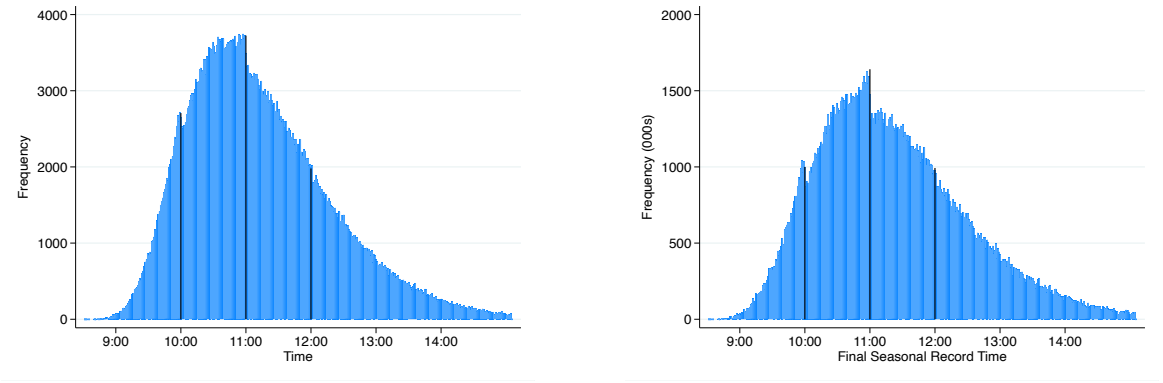


(a) Finishing times

(b) End-of-season record times

Notes: Distribution of finishing times and end-of-season records for the boys 800-meter race in the final sample, 1-second bins. The black line marks the 2:00 reference point.

Figure A9: Boys 3200-meter race: distribution of finishing times and seasonal records

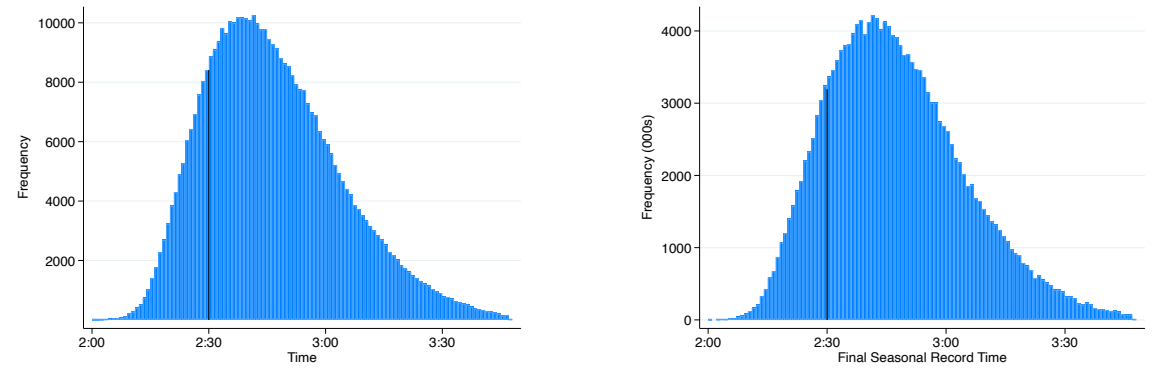


(a) Finishing times

(b) End-of-season record times

Notes: Distribution of finishing times and end-of-season records for the boys 3200-meter race in the final sample, 1-second bins. The black line marks the 10:00 reference point.

Figure A10: Girls 800-meter race: distribution of finishing times and seasonal records

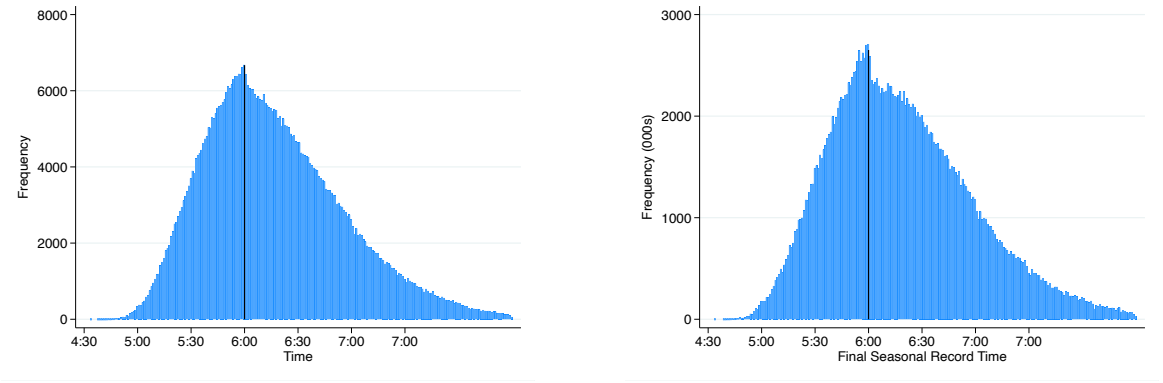


(a) Finishing times

(b) End-of-season record times

Notes: Distribution of finishing times and end-of-season records for the girls 800-meter race in the final sample, 1-second bins. The black line marks the 2:30 reference point.

Figure A11: Girls 1600-meter race: distribution of finishing times and seasonal records

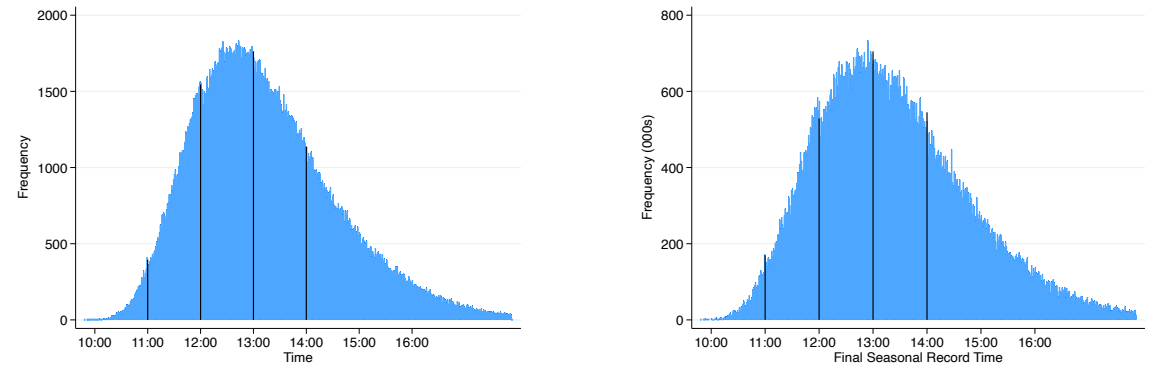


(a) Finishing times

(b) End-of-season record times

Notes: Distribution of finishing times and end-of-season records for the girls 1600-meter race in the final sample, 1-second bins. The black line marks the 6:00 reference point.

Figure A12: Girls 3200-meter race: distribution of finishing times and seasonal records



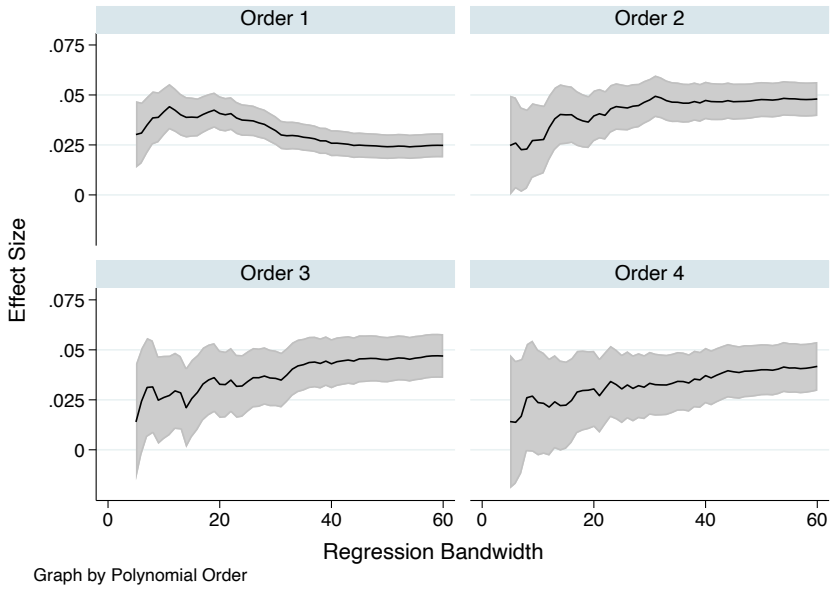
(a) Finishing times

(b) End-of-season record times

Notes: Distribution of finishing times and end-of-season records for the girls 3200-meter race in the final sample, 1-second bins. The black line marks the 11:00 reference point.

A4 Improvement and persistence

Figure A13: Discontinuity in improvement, boys 1600-meter race: robustness to polynomial order and bandwidth



Notes: The discontinuity in the probability of improvement in the next race at 5 minutes, re-estimated across polynomial orders 1–4 and regression bandwidths of 5–60 seconds.

Table A7: Robustness: discontinuity in improvement and participation for different RD specifications: boys 1600-meter race, 5 minutes

	Main	Main + Controls	Main + Athlete FE	Main + Athlete + Race FE	Main +All FE + Controls	Bias-Adjusted	Bias-Adjusted + Controls
Fraction Improving: Next Race	0.04***	0.042***	0.042***	0.041***	0.042***	0.039***	0.041***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Time Improvement: Next Race	-0.36***	-0.4***	-0.579***	-0.57***	-0.57***	-0.329***	-0.376***
	(0.103)	(0.099)	(0.098)	(0.097)	(0.096)	(0.112)	(0.109)
Fraction Improving: Rest of Season	0.034***	0.036***	0.037***	0.037***	0.037***	0.035***	0.037***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Time Improvement: to End of Season Best	-0.362***	-0.423***	-0.581***	-0.598***	-0.593***	-0.326***	-0.392***
	(0.103)	(0.097)	(0.091)	(0.084)	(0.082)	(0.114)	(0.1)
Participation: Any Other Race that Season	0.02***	0.021***	0.023***	0.023***	0.022***	0.016***	0.018***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)
Participation: Number of Races More	0.056***	0.058***	0.061***	0.046***	0.046***	0.044**	0.048**
	(0.019)	(0.018)	(0.017)	(0.015)	(0.015)	(0.02)	(0.019)
Race-specific controls		X			X		X
Athlete-specific controls		X	N/A	N/A	N/A		X
Athlete Fixed Effects			X	X	X		
Race Fixed Effects				X	X		X
Bias Adjustment						X	X

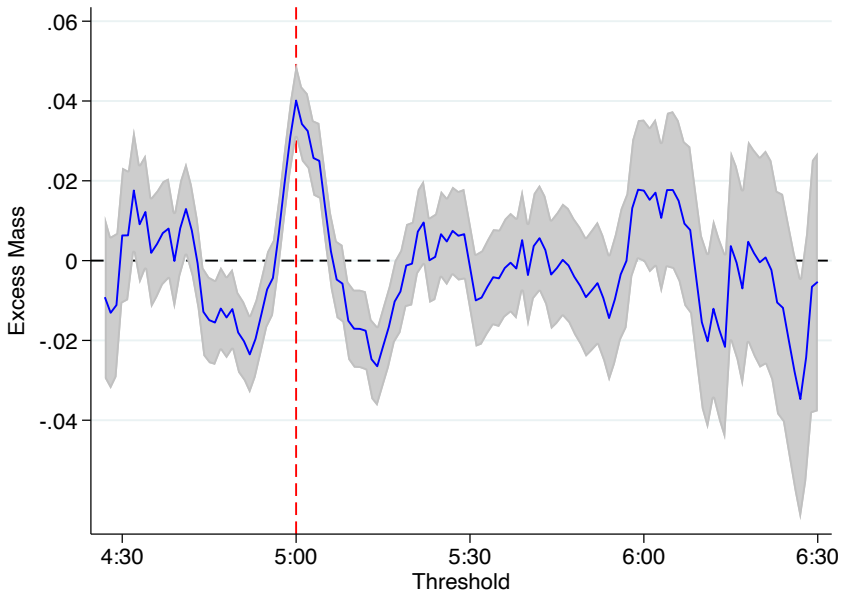
Notes: The discontinuity in each outcome at 5 minutes under a linear RD, progressively adding athlete-race controls and race and athlete fixed effects, and using the Calonico et al. (2014) bias-correction procedure with and without controls. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A8: Robustness: discontinuity in improvement for different exclusion criteria: boys 1600-meter race, 5 minutes

	Effect Size	t-statistic	n (sample size within regression window)
Final Sample	0.040	8.51	161,831
Final Sample, and Including:			
Athletes with hand recorded times:	0.042	13.28	362,166
Athletes not in High School	0.041	8.93	168,374
Athletes with missing times	0.037	8.62	203,000
Athletes with time record issues	0.041	8.58	160,422
Athletes who ran the Mile	0.043	9.13	167,942
Seasons prior to 2009	0.038	7.63	146,883
Athletes with outlier times	0.041	8.60	162,756
Athletes with duplicate time entries	0.044	10.59	210,447
All Athletes	0.041	12.34	342,386

Notes: The discontinuity in the probability of improvement in the next race at 5 minutes, re-estimated on the full (final-plus-excluded) sample and after separately relaxing each exclusion criterion in Table A1, one at a time.

Figure A14: Improvement discontinuity "placebo test": the discontinuity effect size at all threshold times, boys 1600-meter race



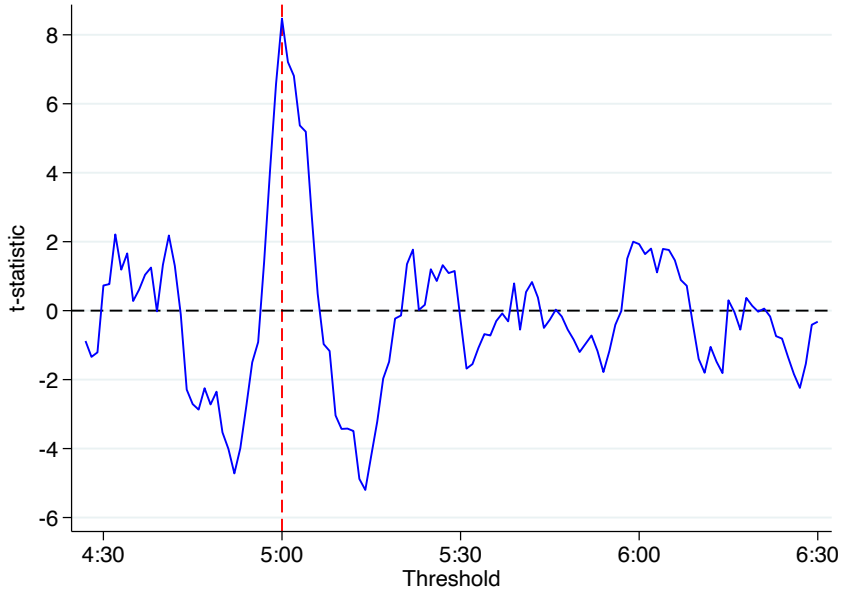
Notes: The estimated discontinuity in the probability of improvement in the next race, re-estimated at every feasible one-second threshold between 4:27 and 6:46.

Table A9: Moderation: discontinuity in improvement for different sub-samples: boys 1600-meter race, 5 minutes

	Effect Size	t-statistic	n (sample size within regression window)
Year			
2009	0.097	2.39	2,266
2010	-0.003	-0.12	4,563
2011	0.037	1.50	6,033
2012	0.056	2.40	6,602
2013	0.060	2.61	7,007
2014	0.035	2.16	13,554
2015	0.074	5.02	16,564
2016	0.014	0.94	16,536
2017	0.054	4.35	24,000
2018	0.024	2.02	25,857
2019	0.040	2.62	16,282
Month			
1	0.027	1.21	6,488
2	0.027	1.49	10,242
3	0.038	3.70	35,835
4	0.039	4.37	45,909
5	0.072	4.94	17,978
Grade			
9	0.000	0.01	14,930
10	0.052	4.90	33,194
11	0.053	4.85	31,321
12	0.036	3.14	29,118
Race Number			
1	0.027	3.89	73,649
2	0.047	4.61	35,845
3	0.049	3.63	19,890
4	0.083	4.15	9,067

Notes: The discontinuity in the probability of improvement in the next race at 5 minutes, re-estimated separately by race year (2009–2019), race month, athlete grade (9th–12th), and race number within the season.

Figure A15: Improvement discontinuity "placebo test":
the discontinuity t-statistic at all threshold times, boys 1600-meter race

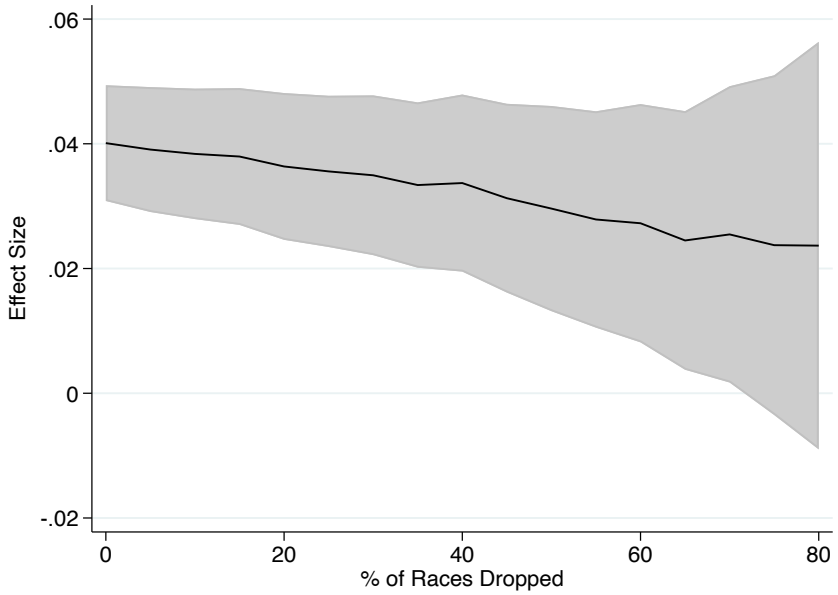


Notes: t-statistics from the same placebo procedure as Figure A14.

A4.1 Robustness to missing data

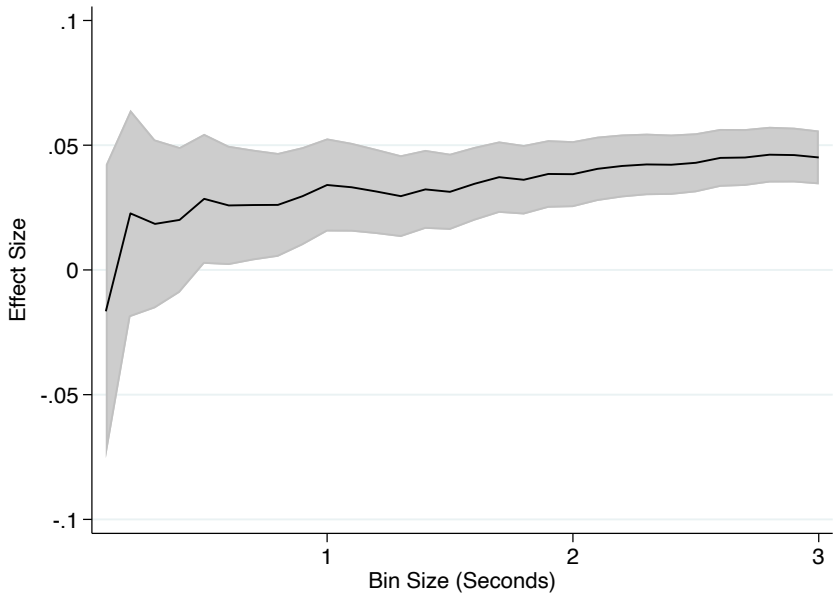
We show that missing data mechanically reduce the size of the discontinuity estimates presented. First, we simulate missing data by randomly dropping observations from athletes' records in our data and estimating the size of the discontinuity using the same RD approach. Simulations show that as data missingness increases, estimates become smaller and noisier (Figure A16). Second, we use two empirical results in this paper to argue that that data missingness should reduce the effects observed. We find that the negative performance effect of surpassing 5 minutes is weaker: i) for improvement in races completed after the next race an athlete runs (i.e., two or more races into the future; see Section 4.3.1), and ii) after performances that are not seasonal records (See Section 4.4.1). If data for a subset of races in an athlete's career are missing, this implies that several observations in our main analysis (which restricts the analysis to SRs), will not be actual SRs, when the missing races include faster times that are not observed. This inclusion of non-SRs in our analysis would suppress the effect estimate observed given the result in Section 4.4.1. Missing races also imply that for several observations in our main analysis, when we are missing the next race after a SR, we determine improvement in the next race (the main dependent variable) using races that are actually two or more races into the future. Given the result in Section 4.3.1, that we observe a smaller discontinuity for improvement in these races, this implies that including these observations in our analysis would suppress the observed discontinuity. Both empirical results thus suggest that missing data would attenuate the observed effect.

Figure A16: Bootstrap estimated improvement discontinuity:
simulations after dropping random observations, by share, boys 1600-meter race



Notes: The discontinuity in the probability of improvement in the next race at 5 minutes, re-estimated after randomly dropping an increasing share of observations from athletes' records to simulate missing data.

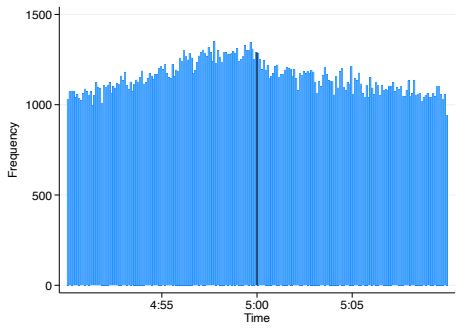
Figure A17: Difference (t-tests) in improvement between times above and below 5 minutes
by the size of the bin/area either side, boys 1600-meter race



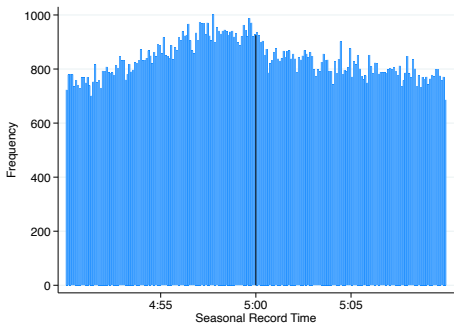
Notes: t-tests of the difference in improvement probability between finishing times just above versus just below 5 minutes, for narrow bins immediately either side of the threshold.

Figure A18: Time and seasonal record distributions at different bin sizes: boys 1600-meter race

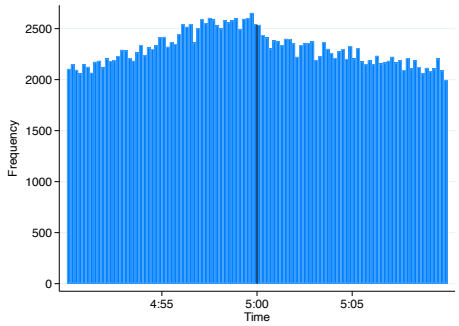
(a) Times: 0.1-second bins



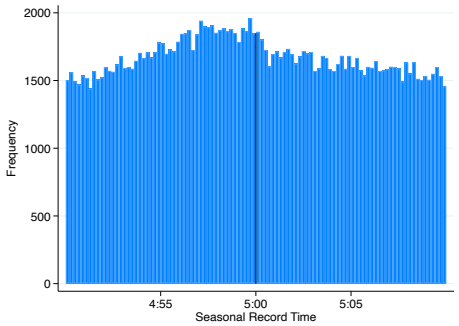
(b) SRs: 0.1-second bins



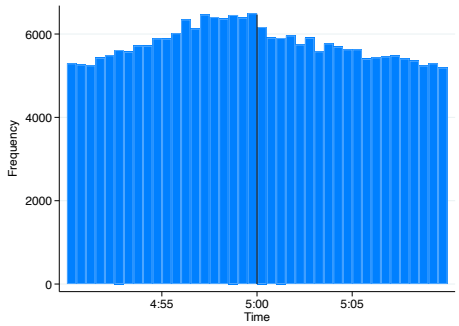
(c) Times: 0.2-second bins



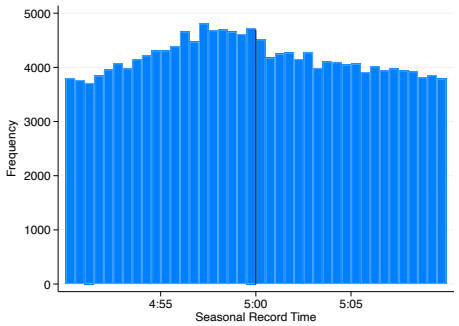
(d) SRs: 0.2-second bins



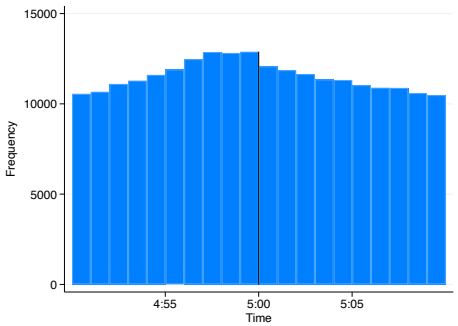
(e) Times: 0.5-second bins



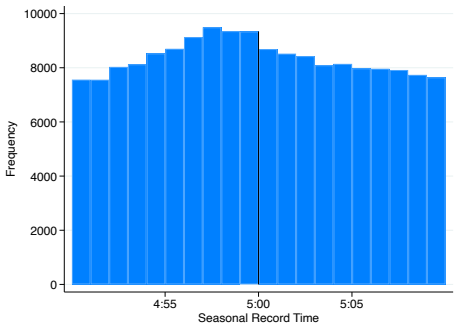
(f) SRs: 0.5-second bins



(g) Times: 1-second bins

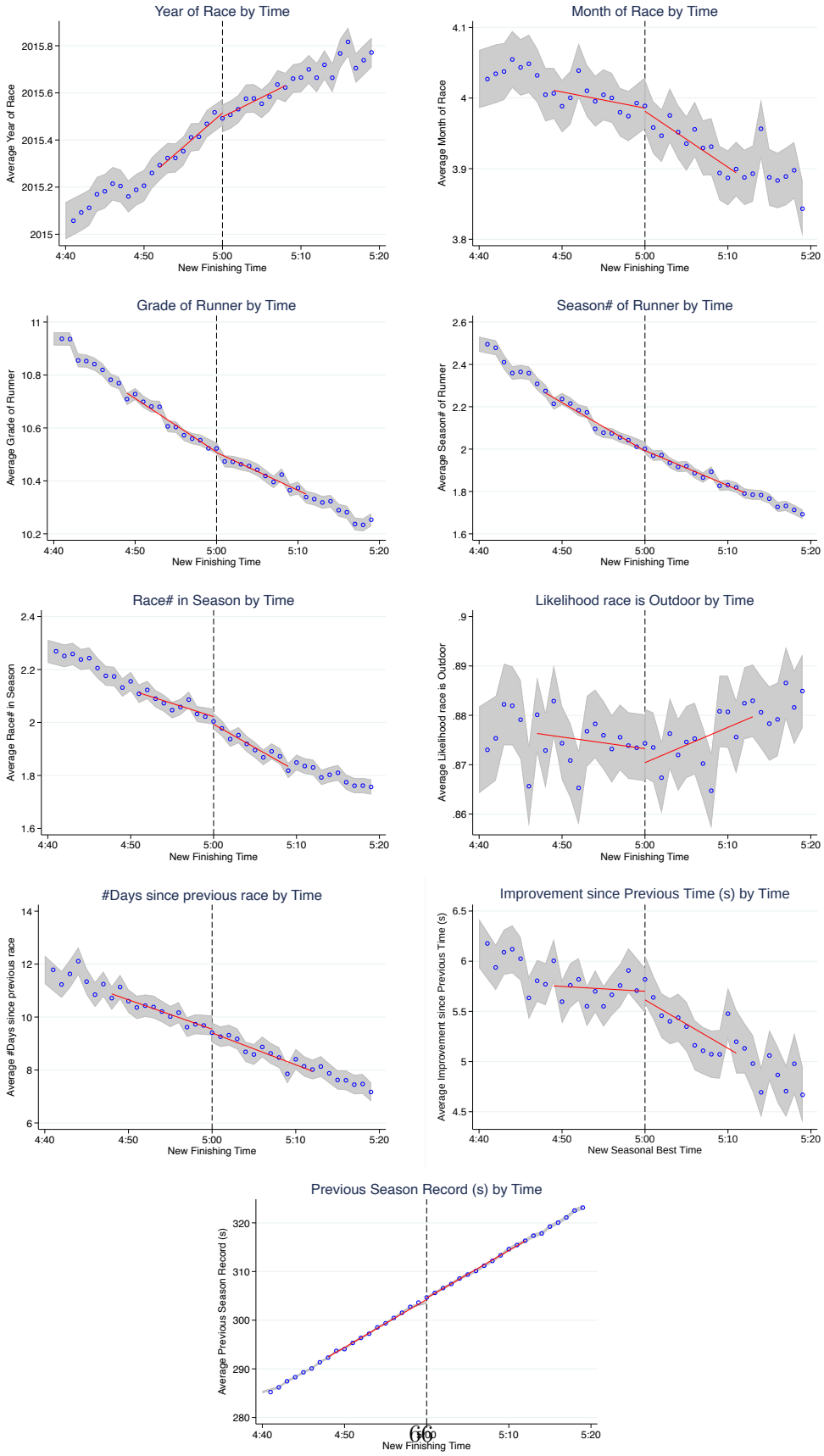


(h) SRs: 1-second bins



Notes: Distribution of finishing times and seasonal records around 5 minutes in the final sample, at progressively coarser bin widths (0.1, 0.2, 0.5, and 1 second).

Figure A19: Distributions and RD plots for other observables as a function of seasonal best in the boys 1600-meter race



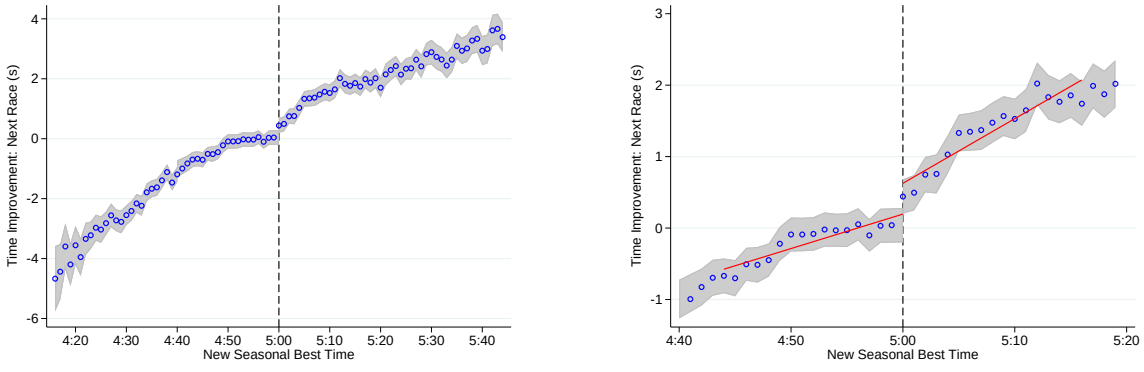
Notes: Binned means and RD fits for the observables in Table A10, in order: race year, month, grade, season number, race number within season, outdoor, days since previous race, improvement since previous time, and previous season's record.

Table A10: Tests for discontinuities in other observables: boys 1600-meter race, 5 minutes

	Main	Bias-Adjusted
Year of Race	-0.029	-0.035
	(0.029)	(0.031)
Month of Race	0.001	0
	(0.015)	(0.017)
Grade of Runner	0.005	0.003
	(0.009)	(0.01)
Season Number of Runner	0.008	0.007
	(0.01)	(0.011)
Race Number in Season	-0.018	-0.021
	(0.014)	(0.016)
Likelihood race is Outdoor	-0.004	-0.002
	(0.003)	(0.003)
Number of days since previous race	1.106	-0.859
	(1.875)	(2.07)
Change since Previous Time (s)	0.043	-0.006
	(0.093)	(0.104)
Previous Record Time (s)	-0.023	0.008
	(0.114)	(0.123)

Notes: Discontinuities at 5 minutes in observable characteristics that should not, under the identifying assumption, jump at the threshold: race year, month, athlete grade, season number, race number within season, whether the race was outdoor, days since the previous race, improvement since the previous time, and the previous season's record. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

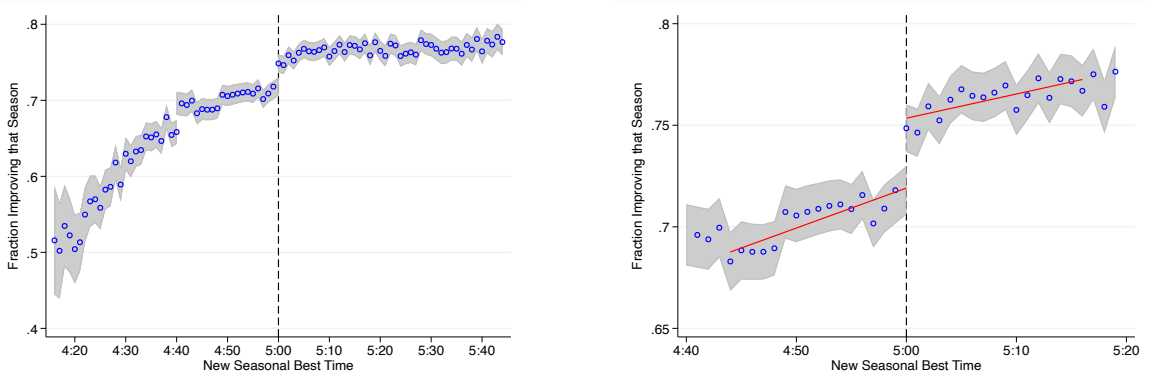
Figure A20: Change in finishing time in the next race: boys 1600-meter race



(a) (b)

Notes: Binned mean improvement in finishing time to the next race, by seasonal record, in the final sample. Panel (a) shows the full range around 5 minutes; panel (b) zooms in on the discontinuity, with separate linear fits on each side.

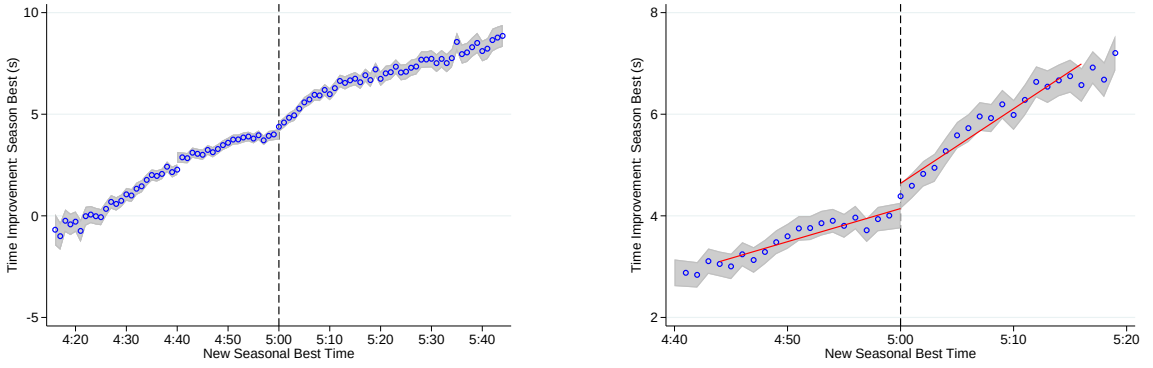
Figure A21: Probability of improvement by the end of the season: boys 1600-meter race



(a) (b)

Notes: Binned mean probability of improving by the end of the season, by seasonal record, in the final sample. Panel (a) shows the full range around 5 minutes; panel (b) zooms in on the discontinuity, with separate linear fits on each side.

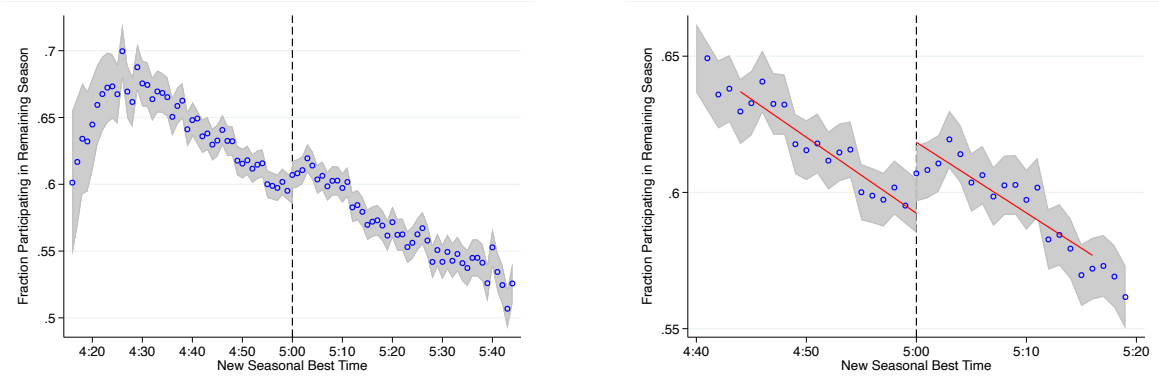
Figure A22: Change in seasonal record time by the end of the season: boys 1600-meter race



(a) (b)

Notes: Binned mean improvement in time by the end of the season (relative to the fastest remaining race that season), by seasonal record, in the final sample. Panel (a) shows the full range around 5 minutes; panel (b) zooms in on the discontinuity, with separate linear fits on each side.

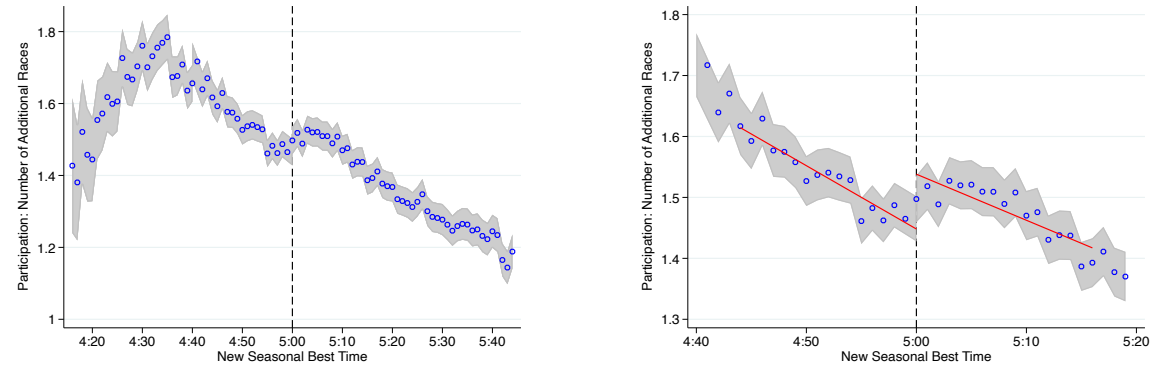
Figure A23: Participation: probability of participating in any other race, boys 1600-meter race



(a) (b)

Notes: Binned mean probability of running another race that season, by seasonal record, in the final sample. Panel (a) shows the full range around 5 minutes; panel (b) zooms in on the discontinuity, with separate linear fits on each side.

Figure A24: Participation: number of additional races, boys 1600-meter race



(a) (b)

Notes: Binned mean number of additional races run that season, by seasonal record, in the final sample. Panel (a) shows the full range around 5 minutes; panel (b) zooms in on the discontinuity, with separate linear fits on each side.

Table A11: Discontinuity in participation and improvement: all events, boys and girls

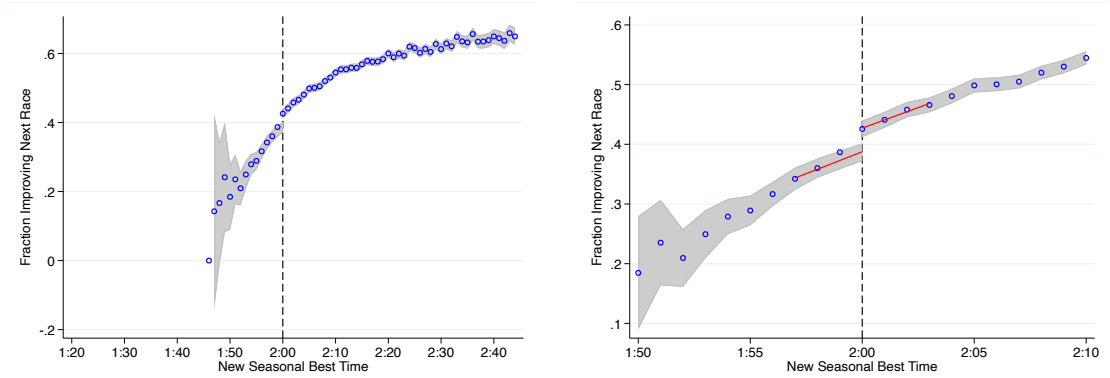
	Boys			Girls		
	800 Meter	1600 Meter	3200 Meter	800 Meter	1600 Meter	3200 Meter
Fraction Improving: Next Race	0.019*	0.04***	0.022***	0.008	0.018***	0.022**
	(0.011)	(0.005)	(0.007)	(0.008)	(0.007)	(0.011)
Time Improvement: Next Race	0.178**	0.36***	0.991***	0.05	0.318*	0.388
	(0.078)	(0.103)	(0.337)	(0.09)	(0.162)	(0.504)
Fraction Improving: Rest of Season	-0.002	0.034***	0.026***	0.016*	0.023***	0.015
	(0.014)	(0.004)	(0.007)	(0.009)	(0.006)	(0.01)
Time Improvement: to End of Season Best	0.052	0.362***	1.014***	0.025	0.189	0.703
	(0.097)	(0.103)	(0.34)	(0.088)	(0.192)	(0.44)
Participation: Number of Races More	0.075	0.056***	0.033*	0.017	0.04*	-0.007
	(0.047)	(0.019)	(0.019)	(0.027)	(0.021)	(0.035)
Participation: Any Other Race that Season	0.008	0.02***	0.012*	0.002	0.013**	0.016*
	(0.01)	(0.005)	(0.006)	(0.006)	(0.006)	(0.009)

Notes: The discontinuity at each event's round-number threshold in the probability of improving in the next race, improvement in time in the next race, the probability of improving by season's end, improvement in time by season's end, the probability of running another race that season, and the number of additional races run, estimated separately for all six event/gender combinations. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure A25: Probability of improvement in the next race:
boys 800-meter and 3200-meter races at their respective thresholds

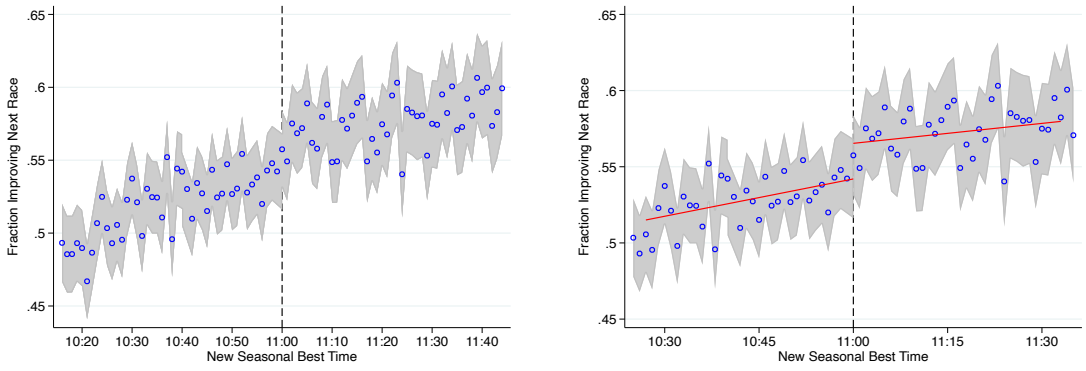
(a) 800-meter race: wide

(b) 800-meter race: zoomed



(c) 3200-meter race: wide

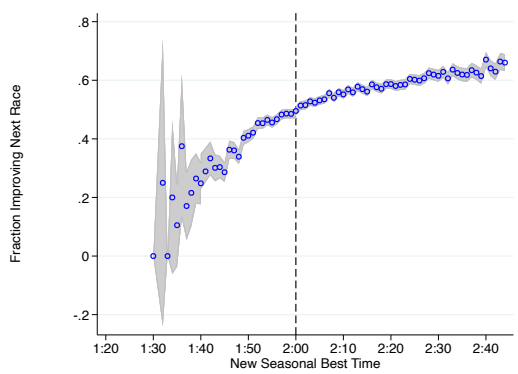
(d) 3200-meter race: zoomed



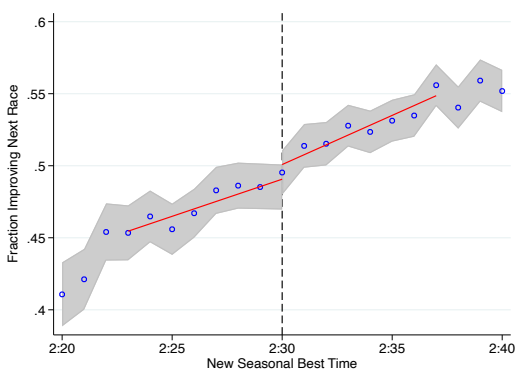
Notes: Binned mean probability of improving in the next race, by seasonal record, in the boys 800-meter (2:00 threshold) and 3200-meter (10:00 threshold) races, in the final sample for each event.

Figure A26: Probability of improvement in the next race:
girls 800-meter, 1600-meter, and 3200-meter races at their respective thresholds

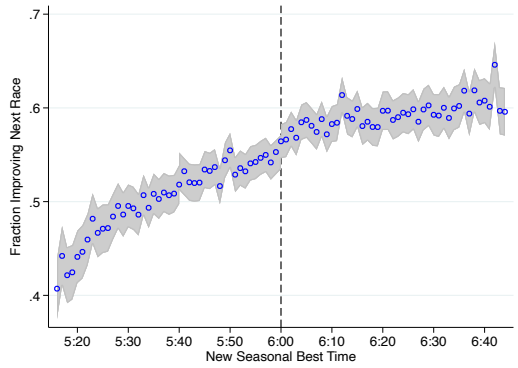
(a) 800-meter race: wide



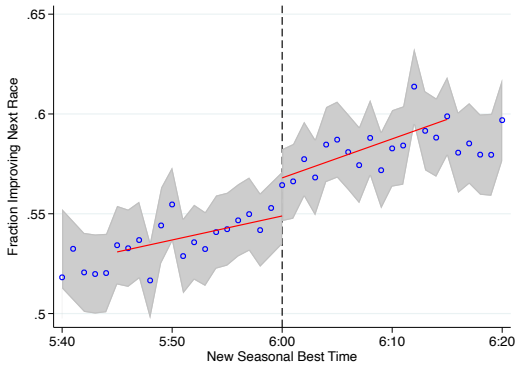
(b) 800-meter race: zoomed



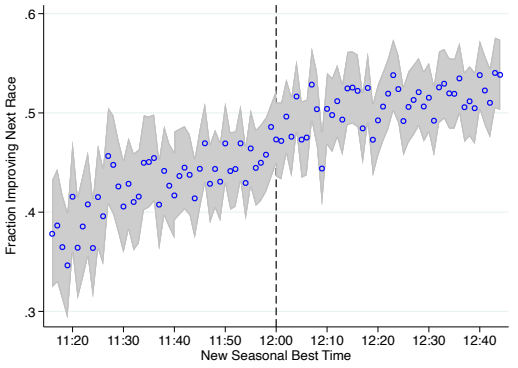
(c) 1600-meter race: wide



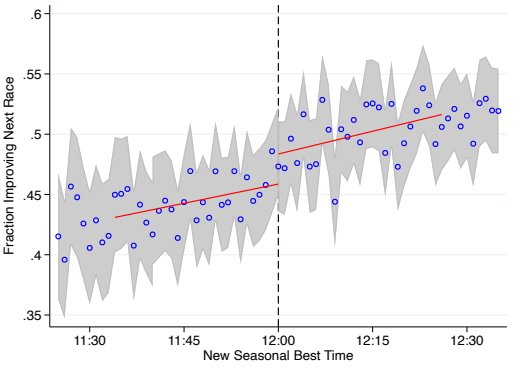
(d) 1600-meter race: zoomed



(e) 3200-meter race: wide



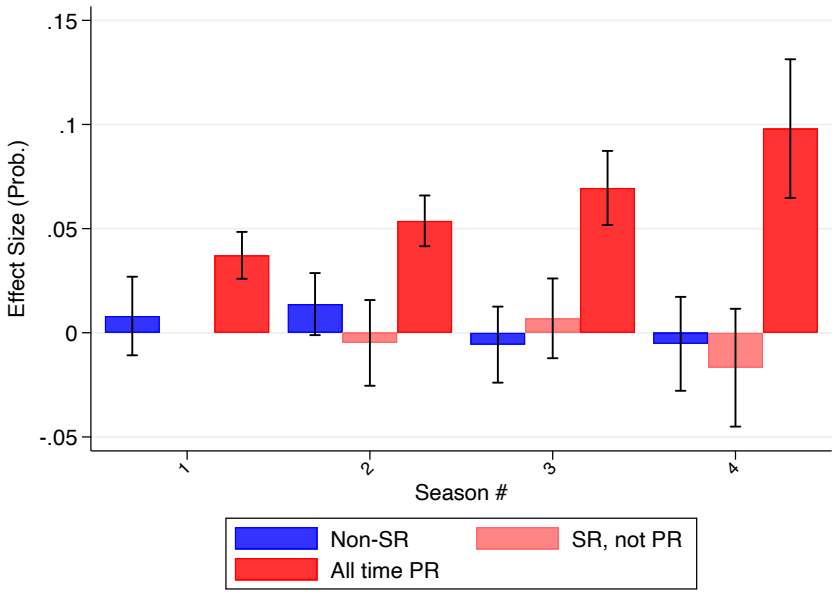
(f) 3200-meter race: zoomed



Notes: Binned mean probability of improving in the next race, by seasonal record, in the girls 800-meter (2:30 threshold), 1600-meter (6:00 threshold), and 3200-meter (11:00 threshold) races, in the final sample for each event.

A5 Improvement mechanisms

Figure A27: Discontinuity in improvement for PRs vs SRs vs non-SRs by athlete season number: boys 1600-meter race



Notes: The discontinuity in the probability of improvement in the next race at 5 minutes, estimated separately by an athlete's season number (1st through 4th) and by whether the seasonal record is also an all-time PR, an SR that is not a PR, or not a seasonal record.

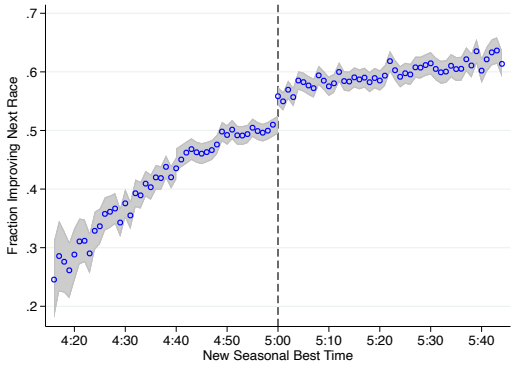
Table A12: Discontinuity in improvement for seasonal records (SRs) compared with non-SR times and different types of non-SR: boys 1600-meter race

	All Seasonal Records (SRs)	Non-SRs	Seasonal Records		Non-SRs: Distance from SR		
			All time PRs	Not PRs	(0s, 3s)	(3s, 7s)	>7s
Fraction Improving: Next Race	0.04***	0.007	0.054***	-0.002	0.021*	-0.011	-0.003
	(0.003)	(0.004)	(0.003)	(0.007)	(0.009)	(0.009)	(0.006)
Time Improvement: Next Race	0.37***	-0.162	0.54***	-0.324	0.095	-0.279	-0.326
	(0.091)	(0.143)	(0.089)	(0.23)	(0.305)	(0.168)	(0.235)
Fraction Improving: Rest of Season	0.045***	0.011**	0.055***	0.005	0.017*	-0.001	0.003
	(0.003)	(0.004)	(0.003)	(0.005)	(0.008)	(0.007)	(0.005)
Time Improvement: to End of Season Best	0.449***	-0.197*	0.536***	-0.095	-0.166	-0.302*	-0.252
	(0.073)	(0.096)	(0.081)	(0.13)	(0.167)	(0.146)	(0.175)
Participation: Number of Races More	0.103***	0.005	0.091***	0.063	0.015	-0.037	0.001
	(0.015)	(0.018)	(0.014)	(0.035)	(0.032)	(0.031)	(0.032)
Participation: Any Other Race that Season	0.024***	0.002	0.026***	0.011*	0.012	-0.015*	0.002
	(0.003)	(0.004)	(0.003)	(0.005)	(0.006)	(0.008)	(0.007)

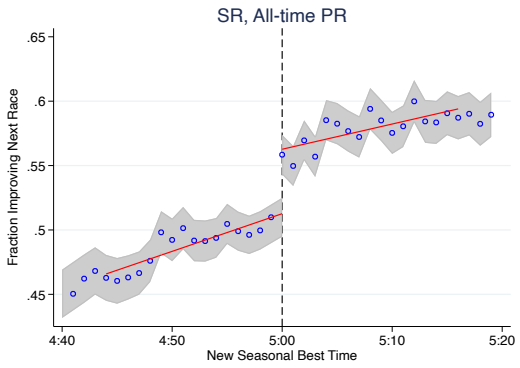
Notes: The discontinuity in the probability of improvement in the next race at 5 minutes, estimated separately for seasonal records versus non-seasonal-record times at increasing distances from the seasonal record (top panel), and for seasonal records that are all-time PRs, seasonal records that are not, and non-seasonal-record times (bottom panel). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure A28: Probability of improvement in the next race by previous sub-5 performances: boys 1600-meter race

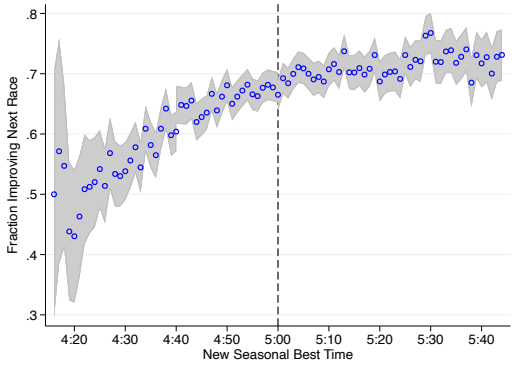
(a) All-time PR: wide



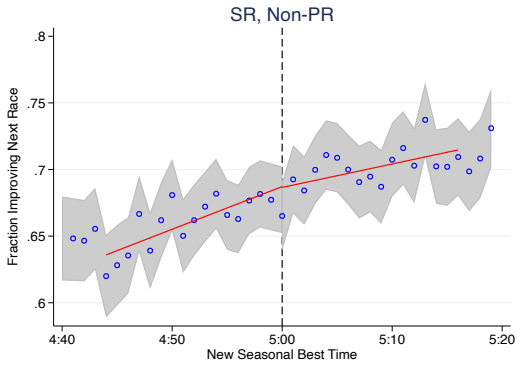
(b) All-time PR: zoomed



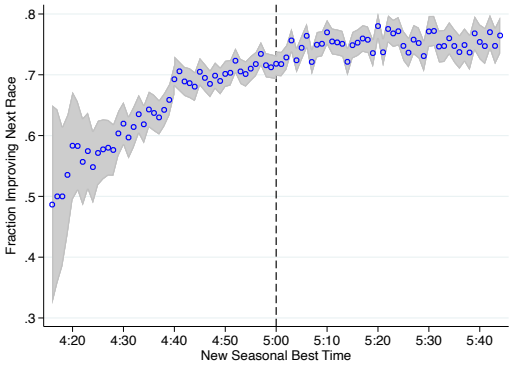
(c) SR, not PR: wide



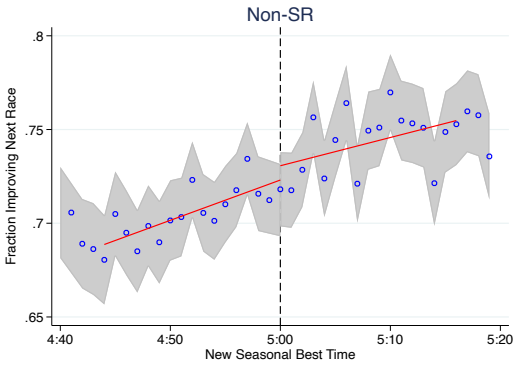
(d) SR, not PR: zoomed



(e) Non-SR: wide

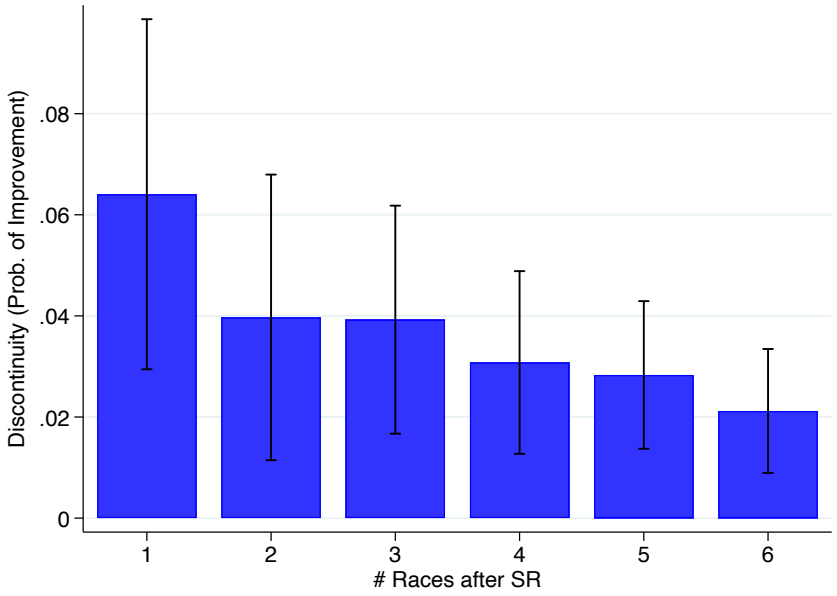


(f) Non-SR: zoomed



Notes: Binned mean probability of improving in the next race, by seasonal record, in the final sample, separately for seasonal records that are all-time PRs, seasonal records that are not, and non-seasonal-record times.

Figure A29: Discontinuity in improvement by n races in the future, conditional on completing 6 races:
boys 1600-meter race



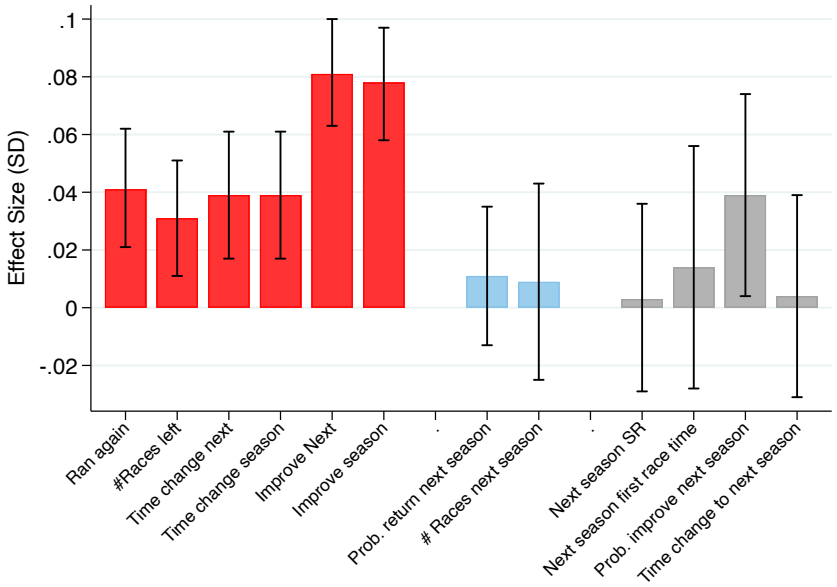
Notes: The discontinuity in the probability of improvement at 5 minutes, estimated separately for each of the six races following a seasonal record, restricted to athletes who ran at least six more races that season.

Table A13: Longer-term effects: discontinuity in participation and improvement next season: boys 1600-meter race, 5 minutes

	Final Sample			All Seasons (without recording errors)		
	Effect Size	t-statistic	n (sample size within regression window)	Effect Size	t-statistic	n (sample size within regression window)
Participation Outcomes						
Prob. return next season	-0.0062	1.06	95,707	-0.0097	2.10	148,787
# Races next season	-0.0239	0.63	50,213	-0.0643	1.85	77,312
Improvement Outcomes						
Next season SR	0.1028	0.45	49,272	-0.0337	0.21	100,771
Next season first race time	-0.1663	0.52	31,802	0.0178	0.07	87,065
Prob. improve next season	0.0178	2.22	48,026	0.0146	2.68	100,948
Time change to next season	0.1028	0.45	49,272	-0.0337	0.21	100,771

Notes: The discontinuity at 5 minutes in whether an athlete returns the following season, next season's number of races, seasonal record, and first-race time, and the probability of and amount of improvement in next season's seasonal record relative to the present season's, restricted to athletes in grades 9–11.

Figure A30: Longer-term effects: discontinuity in participation and improvement next season: boys 1600-meter race, 5 minutes



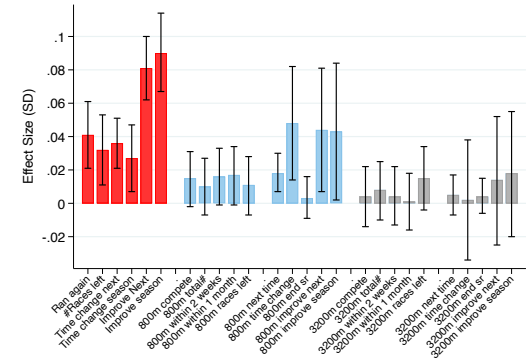
Notes: Effect sizes (in standard deviation units) for the within-season outcomes in Table 2 alongside the next-season outcomes in Table A13, for comparison.

Table A14: Substitution to other events:
discontinuity in participation and improvement for the boys 800-meter and 3200-meter races

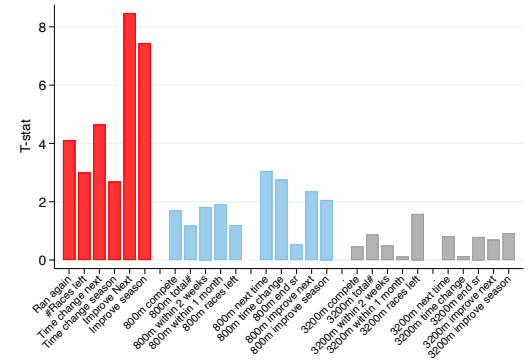
	800 Meter			3200 Meter		
	Effect Size	t-statistic	n (sample size within regression window)	Effect Size	t-statistic	n (sample size within regression window)
Participation Outcomes						
Participate at all	0.0073	1.70	205,970	-0.0021	0.46	188,409
Ran race within 2 weeks	0.0056	1.82	196,896	0.0016	0.49	205,195
Ran race within 1 month	0.0069	1.90	192,534	0.0004	0.12	214,076
Total races	0.0200	1.17	182,276	-0.0152	0.87	193,181
Number of races more	0.0141	1.20	208,555	-0.0203	1.58	199,413
Improvement Outcomes						
Fraction improving: next race	0.0209	2.35	42,444	0.0065	0.69	39,673
Fraction improving: season	0.0164	2.06	35,423	-0.0071	0.92	41,561
Time in next race	-0.2539	3.05	75,297	-0.3121	0.81	66,118
Time improvement next race	-0.2829	2.75	40,900	0.0632	0.13	35,354
End of season record	-0.0454	0.54	88,371	-0.2677	0.79	99,927

Notes: The discontinuity at 5 minutes in the boys 1600-meter race in participation, timing, and improvement outcomes in the 800-meter and 3200-meter races, for athletes who also compete in those events that season.

Figure A31: Substitution: comparing main effects with effects on participation and improvement in the boys 800-meter and 3200-meter races

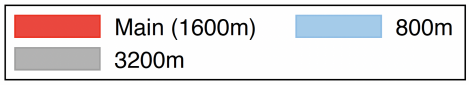


(a) (Absolute) effect size



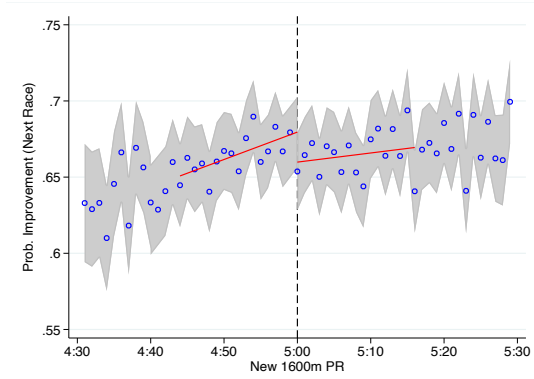
(b) (Absolute) t-statistic

Legend

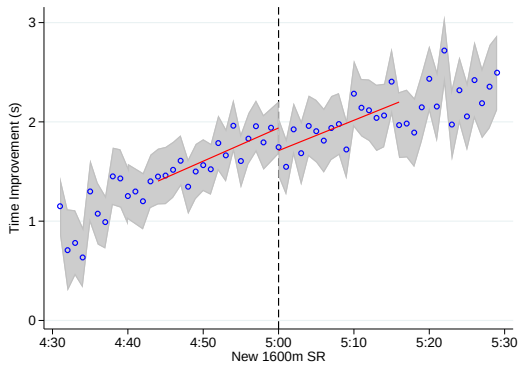


Notes: Effect sizes (in standard deviation units) and t-statistics for the discontinuity at 5 minutes in the boys 1600-meter race, for the within-1600-meter outcomes in Table 2 alongside the 800-meter and 3200-meter substitution outcomes in Table A14, grouped by color.

Figure A32: Substitution: improvement in the boys 800-meter race given boys 1600-meter best times



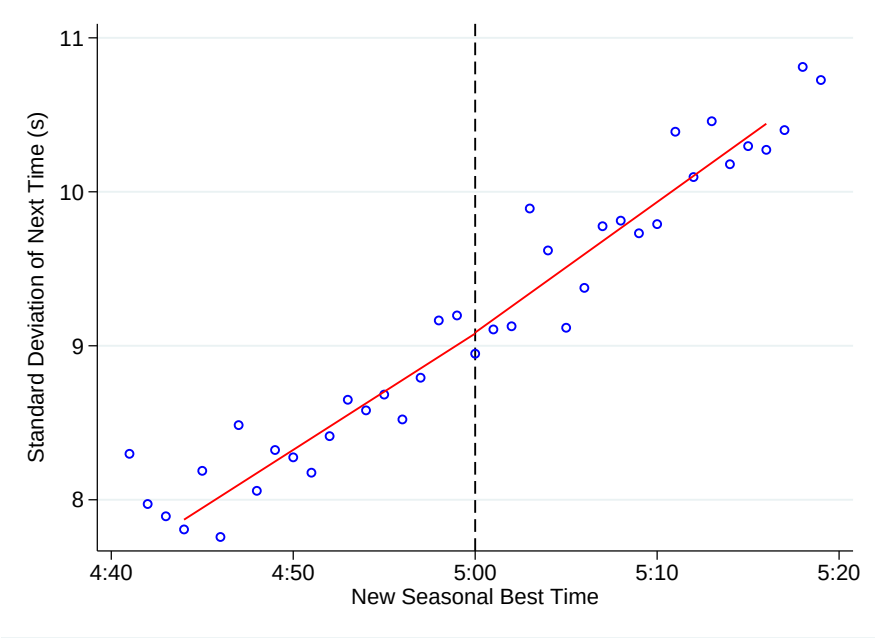
(a) Prob. improvement in next race



(b) Time change in next race

Notes: Binned means and RD fits for the probability of improvement and change in time in the next 800-meter race, as a function of the athlete's boys 1600-meter seasonal record, for athletes who compete in both events.

Figure A33: Standard deviation of finishing times given new seasonal record times: boys 1600-meter race



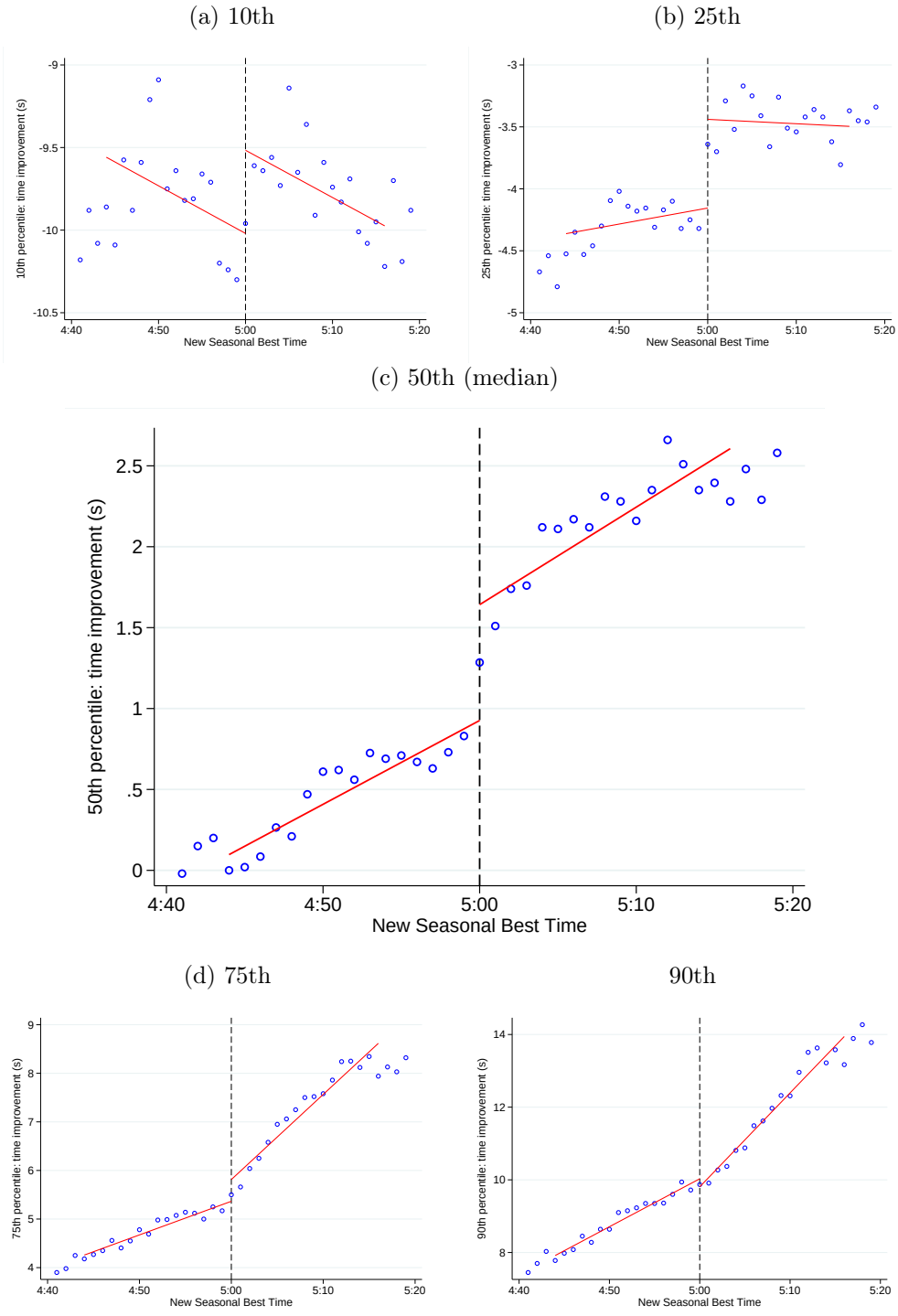
Notes: Binned standard deviation of the next race's finishing time, by seasonal record, in the final sample, with a linear fit on each side of 5 minutes.

Table A15: Discontinuity in improvement for each quantile, using quantile regression: boys 1600-meter race, 5 minutes

	10th %tile	20th %tile	30th %tile	40th %tile	50th %tile	60th %tile	70th %tile	80th %tile	90th %tile
Time Change: Next Race	0.624**	0.738***	0.805***	0.732***	0.628***	0.542***	0.353***	-0.039	-0.378**
	(0.207)	(0.145)	(0.12)	(0.104)	(0.097)	(0.095)	(0.101)	(0.114)	(0.147)
Time Improvement: to End of Season	0.829***	0.848***	0.855***	0.696***	0.376***	0.043	-0.201	-0.112	-0.097
	(0.201)	(0.139)	(0.116)	(0.107)	(0.103)	(0.104)	(0.116)	(0.136)	(0.181)

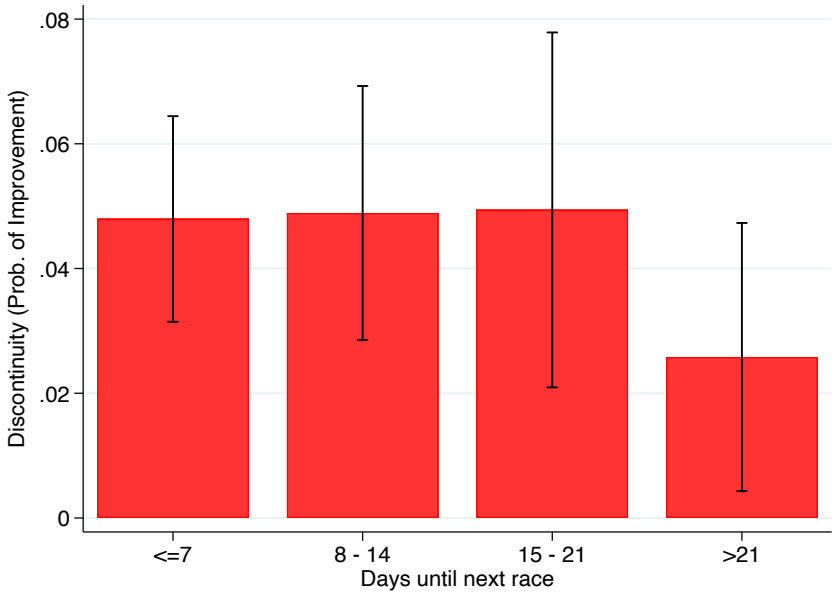
Notes: The discontinuity at 5 minutes in improvement in time to the next race, estimated by quantile regression at the 10th, 25th, 50th, 75th, and 90th percentiles.

Figure A34: Discontinuities in time improvement to the next race at different quantiles of time improvement: boys 1600-meter race



Notes: Binned 10th, 25th, 50th, 75th, and 90th percentiles of improvement in time to the next race, by seasonal record, in the final sample, with a linear fit on each side of 5 minutes.

Figure A35: Discontinuity in improvement in the next race by the # of days until the next race: boys 1600-meter race



Notes: The discontinuity in the probability of improvement at 5 minutes, estimated separately by the number of days until the next race (0–7, 8–14, 15–21, and 22 or more days).

A6 Other reference points

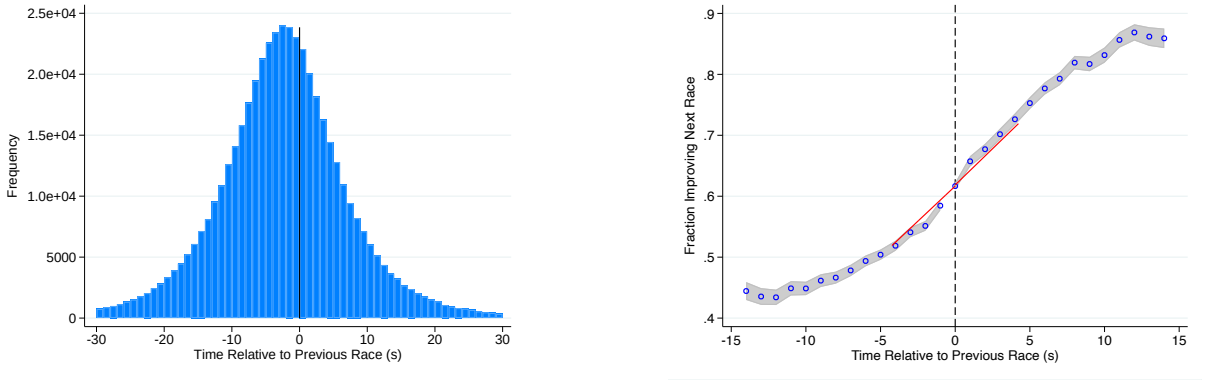
Previous race times or a personal records entering the race could also serve as reference points for runners. Table A16 and Figures A36–A37 repeat the bunching and improvement-discontinuity procedures described above, but re-centered on these alternative reference points: the running variable is an athlete's current time minus their previous race time (Figure A36), or minus their seasonal record entering the race (Figure A37), with the reference point at 0 in both cases. Because these running variables are less diffusely distributed than finishing time around a round number, the bunching region and local window are set to 2 and 10 seconds respectively, rather than the 5- and 25-second values used in the main specification. We find less evidence of reference dependence with respect to both previous race times and seasonal records compared to round numbers, with excess mass estimates of 1.6% and 2.3% respectively, compared to 11.1% for 5 minutes in the boys 1600-meter race. We also do not find discontinuities in improvement in the next race after narrowly surpassing either threshold. These reference points thus appear less important than round numbers in this setting: they generate less bunching and do not produce dynamic effects on improvement.

Table A16: Bunching and improvement discontinuity at alternative reference points: boys 1600-meter race

	Excess Mass				Discontinuity in Improvement Probability		
	Actual finishers	Counterfactual finishers	% excess finishers	t-statistic	Effect Size	t-statistic	n
Previous Time	52,851	52,017	1.6	3.12	0.00	0.34	108,825
Seasonal Record Time	56,051	54,782	2.3	5.05	0.00	-0.57	105,147

Notes: Bunching rows report excess mass, actual and counterfactual finisher counts, and t-statistics from the procedure in Appendix A3.1 (bunching region of 2 seconds, local window of 10 seconds), applied to the running variable at 0. RD rows report the discontinuity in the probability of improvement in the next race at 0, using Calonico et al. (2014)'s optimal bandwidth selector.

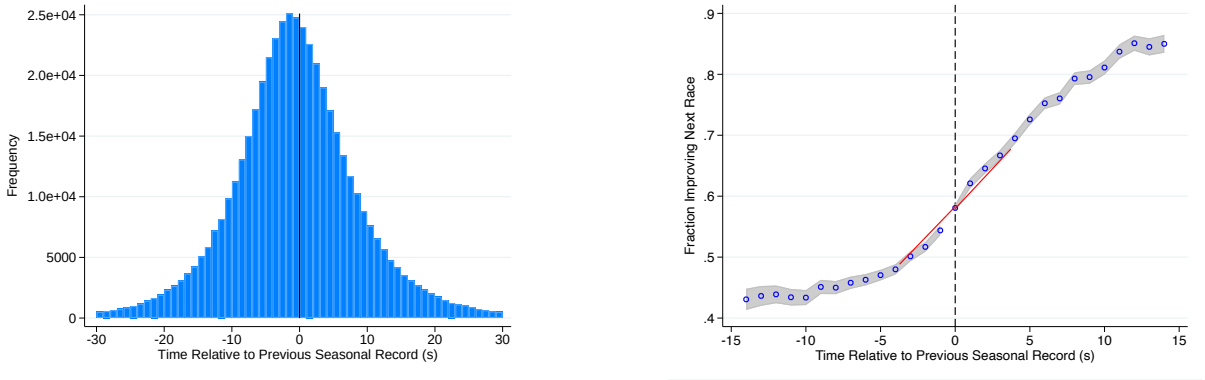
Figure A36: Previous race time as a reference point: boys 1600-meter race



(a) Bunching (b) Improvement discontinuity

Notes: Panel (a) shows the distribution of finishing time minus previous race time in the final sample, 1-second bins; the black line marks 0. Panel (b) shows the binned mean probability of improving in the next race as a function of finishing time minus previous race time, with a fitted line from the RD regression on each side of 0.

Figure A37: Previous seasonal record as a reference point: boys 1600-meter race



(a) Bunching (b) Improvement discontinuity

Notes: Panel (a) shows the distribution of finishing time minus the seasonal record entering the race, in the final sample, 1-second bins; the black line marks 0. Panel (b) shows the binned mean probability of improving in the next race as a function of finishing time minus the seasonal record entering the race, with a fitted line from the RD regression on each side of 0.

A7 Dynamic reference dependence model

A7.1 Model setup

We simulate race times for a population of N runners observed over three periods, $p \in \{1, 2, 3\}$. For each period, we follow a similar set-up to Allen et al (2017), where each runner i chooses a time t_{ip} in each period p to maximize their utility from the performance in that period. This time is defined as the number of seconds faster than the slowest time in this setting; as such, faster performances are increasing in t_{ip}. Each athlete faces a value function $V(t)$ and a cost function $C(t)$ where both value and costs of performances are increasing in t_{ip}. The equilibrium time in each period is determined through myopic optimization, such that a runner selects a performance (and corresponding time t_{ip}) on the first instance that marginal cost equals marginal benefit, i.e., where $C'(t) = V'(t)$. Our base value function is a piecewise linear model described in Section A7.1.1, updated as described in Section A7.1.4. We also consider two alternative value functions – a probabilistic marginal benefit and a diminishing-sensitivity value function – as robustness checks in Section **??**.

A7.1.1 Value function

Runners have a piecewise linear value function around a reference point r, reflecting reference dependent preferences, where r is a 5-minute time (in this model, defined as 110 seconds, as it is this amount faster than the slowest time, 6 minutes 50 seconds).

$$V_{ip}(t) = \begin{cases} t - \eta_{ip}\lambda(r - t), & t \leq r, \\ t + \eta_{ip}(t - r), & t > r. \end{cases}$$

The marginal value (benefit) of time is thus piecewise constant around r.

$$V'_{ip}(t) = \begin{cases} 1 + \eta_{ip}\lambda, & t \leq r, \\ 1 + \eta_{ip}, & t > r. \end{cases}$$

The parameter λ reflects loss aversion, and thus the size of the discontinuity in marginal benefit at the reference point, while η_{ip} defines the strength of reference dependent preferences.

We use a piecewise linear value function in place of a model with diminishing sensitivity because both models perform similarly (see Section A7.3.2) and the piecewise linear model is more parsimonious, easier to fit, and a commonly used simplification of such models.

A7.1.2 Marginal cost

The marginal cost of achieving time t_{ip} is a power function given by

$$C'_{ip}(t) = A_{ip} t_{ip}^{\beta},$$

where A_{ip} scales the cost of effort for runner i in period p (call it their "ability") and $\beta > 0$ controls the curvature of marginal cost, with higher values reflecting steeper marginal costs (and thus lower sensitivity to the value function when determining times).

We express each runner's ability in units of time, as the time they would run under a standard linear value function (i.e., in the absence of reference dependence, where $V_{ip}(t) = t_{ip}$). We call this their "benchmark time", t^*_{ip}. Under a linear value function the equilibrium condition is $A_{ip} t_{ip}^{\beta} = 1$, so that

$$A_{ip} = \frac{1}{(t^*_{ip})^{\beta}},$$

We set benchmark times in period 1 to be uniformly distributed, $t^*_{i1} \sim U(60, 140)$, with the reference point $r = 110$.

Change in times between periods To simulate time changes between periods, we draw on properties of time changes in the empirical data. Times between periods improve slightly on average (i.e., runners get faster), but with relatively high variance (i.e., time change is noisy) and are approximately normally distributed. Changes are both persistent and negatively auto-correlated, i.e., runners who improve by more in period 2 are on average both faster in period 3 and improve by less between periods 2 and 3, consistent with stochasticity in times and regression to the mean. We thus decompose each runner's benchmark time into an underlying fitness level (persistence) and period-specific noise,

$$t^*_{ip} = \theta_{ip} + \varepsilon_{ip}, \qquad \theta_{ip} = \theta_{i,p-1} + u_{ip}, \qquad u_{ip} \sim \mathcal{N}(\mu, \sigma_u^2), \quad \varepsilon_{ip} \sim \mathcal{N}(0, \sigma_\varepsilon^2),$$

where fitness θ_{ip} accumulates across periods and noise ε_{ip} is drawn afresh each period, capturing period-specific variation in form, conditions, or competition.[17]

Following properties of the empirical data, we model time changes $\Delta t^* \sim \mathcal{N}(1, 36)$, with a lag-one autocorrelation $\rho = -0.4$.[18] These two inputs pin down the variances of fitness changes and period-specific noise. From the following

$$\mathrm{Var}(\Delta t^*) = \sigma_u^2 + 2\sigma_\varepsilon^2, \qquad \mathrm{Cov}(\Delta t_p^*, \Delta t_{p+1}^*) = -\sigma_\varepsilon^2.$$

, with $\rho = \mathrm{Cov}/\mathrm{Var}$ gives $\sigma_\varepsilon^2 = -\rho\,\mathrm{Var}(\Delta t^*)$ and $\sigma_u^2 = \mathrm{Var}(\Delta t^*)(1+2\rho)$, we establish $\sigma_\varepsilon = 3.79$ and $\sigma_u = 2.68$. Period-specific idiosyncratic factors thus account for the bulk of the variance in time changes between periods, reflecting that time change between each period is noisy.

A7.1.3 Myopic equilibrium rule

In each period, runners choose t_{ip} by solving $V'_{ip}(t) = C'_{ip}(t)$. The solution is computed sequentially across the two regions of the marginal benefit function for the case of reference dependent preferences, following a "myopic optimization" rule (i.e., the first point where marginal costs equal marginal benefits).

First, the model solves the lower-region equation:

$$A_{ip} t^\beta = 1 + \eta_{ip}\lambda,$$

which gives

$$t_{ip}^L = \left(\frac{1 + \eta_{ip}\lambda}{A_{ip}}\right)^{1/\beta}.$$

If $t_{ip}^L < r$, the runner is assigned

[17] In period 1 we set $\theta_{i1} = t_{i1}^* - \varepsilon_{i1}$, treating the calibrated period-1 benchmark as already containing a transitory component, which preserves the period-1 distribution exactly and gives every period the same noise structure.

[18] These are coarse approximations from the empirical data. ρ is bounded below by -0.5, the limiting case in which no change in ability is persistent and every apparent improvement is given back, and above by 0, a pure random walk in which nothing reverts.

$$t_{ip} = t_{ip}^{L}.$$

If $t_{ip}^{L} \geq r$, the model instead solves the upper-region equation:

$$t_{ip}^{H} = \left(\frac{1 + \eta_{ip}}{A_{ip}} \right)^{1/\beta}.$$

If this upper-region solution lies below the reference point (while the lower region solution lies above), the runner is assigned a value just above the threshold.

$$t_{ip} = r + \varepsilon.$$

Otherwise, the runner is assigned

$$t_{ip} = t_{ip}^{H}.$$

This procedure is applied sequentially for periods $p = 1, 2, 3$, given the changing ability parameters.

A7.1.4 Deflated weight on reference dependence

In the base model of reference dependent preferences, the value function is the same for each participant in each period. We compare results of this model to one where runners who have previously achieved a time above the reference point (i.e., surpassed the reference point r) reduce the weight placed on reference dependence to η^{PR} in periods following surpassing the reference point. This is tantamount to the reference point becoming less important after surpassing it.

Thus,

$$\eta_{i1} = \eta,$$

$$\eta_{i2} = \begin{cases} \eta^{PR}, & t_{i1} > r, \\ \eta, & t_{i1} \leq r, \end{cases}$$

and

$$\eta_{i3} = \begin{cases} \eta^{PR}, & t_{i1} > r \text{ or } t_{i2} > r, \\ \eta, & t_{i1} \leq r \text{ and } t_{i2} \leq r. \end{cases}$$

where

$$\eta^{PR} \leq \eta$$

A7.1.5 Final setup

Models and simulated moments We simulate the linear model under two rules for updating η between periods: baseline (no updating) and deflated (η reduced to η^{PR} after surpassing r). From each simulated dataset we compute: (i) the Chetty bunching statistic for the distribution of times and personal records (PRs) around the reference point; and (ii) OLS regression-discontinuity estimates for the jump in time improvement and improvement probability at the reference point, following both times and new PRs (the latter restricted to runners who set a new PR). Two alternative value functions – a probabilistic marginal benefit and a diminishing-sensitivity specification – are considered as robustness checks in Appendix **??**, since neither outperforms the linear model and both add at least one free parameter.

Parameter values We use $\lambda = 2$, $\eta = 0.1$, $\beta = 5$, $\eta^{PR} = 0.035$ for our base specifications, but also (1) conduct sensitivity analyses to each of these parameters, and (2) use simulated method of moments to fit the values of λ, η and η^{PR} that best produce key moments in our empirical data. The value $\lambda = 2$ follows other literature on reference dependence [cites], the value $\eta = 0.1$ follows calibration from Allen et al. (2017; though in their work, instead of a weight on reference dependence they apply such preferences to a random share of the population, similarly modulating

the effect of reference dependent preferences in aggregate), and $\beta = 5$ implies that a runner will run 4 seconds faster when sufficiently close to surpassing the reference point compared to a model with no reference dependence (a somewhat arbitrary, but potentially reasonable assumption for this set of competitive athletes).

A7.2 Simulation results

A7.2.1 Results summary

First, while the standard ("baseline") model of reference dependent preferences generates bunching in personal record times and finishing times, it fails to produce the discontinuities in improvement observed in our empirical data. For the discontinuity in improvement in finishing times, this model generates the opposite of the empirical result: runners narrowly surpassing 5 minutes improve their times significantly *more* relative to those narrowly missing 5 minutes. For the discontinuity in improvement probability, the standard model produces no discontinuity at the reference point whatsoever.

For both discontinuities, the "deflated model", where the weight on reference dependence attenuates after it has been surpassed, produces the expected negative effect in time improvement and improvement probability. Moreover, while the baseline model produces bunching, the deflated model produces more bunching in PRs than in times, as we observe in the empirical data. Figures A42-A44 show example plots illustrating these patterns, and Table A17 summarizes these results quantitatively for both fixed and fitted model parameters, showing the same patterns described above. The deflated model also performs better on measures of model fit, described in Section A7.2.2.

Figure A38 illustrates the mechanics of the deviations between simulated results for each model. It compares two runners who in period 1 fall either just short of the reference point (5:02; left-hand panel) or just past it (4:58; right-hand panel). Each then faces the same family of period-2 ability changes, illustrated by cost curves c'_{2a} through c'_{2g}, centered on the mean change (a slight improvement in ability) derived from the empirical data.

Whether a given shock produces an improvement or a decline is largely a property of the shock itself, not of reference dependence or the value function: four of the seven shocks improve the near-

miss runner's time, and the same four improve the narrow-gain runner's time under the baseline model – an identical count, because both runners face the same (undeflated) marginal-benefit curve in period 2. The sign of the ability change thus decides the sign of time change under the baseline model (i.e., improvement probability). This is exactly the claim above that improvement probability should not jump discontinuously at r under the baseline model. Deflation changes this: the same seven shocks improve the narrow-gain runner only three times out of seven, because the mean shock ($\delta = 1$, curve c'_{2d}) – which moves the standard model runner from 4:58 to 4:57, an improvement – instead leaves the deflated runner at 4:58.04, a very small decline. The same ability change for both models flips from a gain to a (tiny) loss because η^{PR} replaces η in period 2. Integrating over the model's full distribution of ability changes rather than these seven, the near-miss runner improves with probability 56.6%, as does the narrow-gain runner under the baseline model, while the deflated narrow-gain runner improves with probability 49.8%—a negative discontinuity of 6.8 percentage points. This basic example illustrates how the deflated model generates a negative discontinuity in the probability of improvement.

The same exercise explains the time-improvement discontinuity's sign. Averaging the resulting equilibria over the model's full distribution of ability changes, the near-miss runner improves by 0.28 seconds on average, while the narrow-gain runner under the baseline model improves by *more*, 1.50 seconds, the opposite of the empirical data. The individual curves illustrate why. The most negative shock (c'_{2a}, $\delta = -5$), for example, costs the near-miss runner 5.19 seconds but only 3.22 for the baseline narrow-gain runner: because the narrow-gain runner's new equilibrium is below the reference point, the higher marginal benefit arrests their decline as they try to mitigate newly experienced losses. By contrast, the most positive shock (c'_{2g}, $\delta = 7$) improves each by 5.27 and 7.13 seconds respectively—now the near-miss runner moves past the reference point, and the reduced marginal benefit arrests their improvement, as they slack off after surpassing 5 minutes. In sum, narrow-gain runners lose less for reductions in ability and gain more for improvements. Under the deflated weight on reference dependence, the narrow-gain runner's average improvement falls to 0.31 seconds, close to the near-miss runner's 0.28, because the lower η^{PR} reduces motivation for all changes in ability.

While a toy example, this approximates the distribution of ability changes in the model, and

so illustrates expected changes in finishing times between periods for near misses and narrow wins. This deflated weight thus allows these model predictions to more closely approximate the reductions in improvement following narrow wins in our empirical data. Table A17 reports the key moments quantitatively for each model, for both fixed (calibrated) and SMM-fitted parameter values. Section **??** shows that the deflated model also performs better than a baseline model under alternative functional forms for the value function; specifically, a value function with diminishing sensitivity and one where time realizes with some noise rather than deterministically.

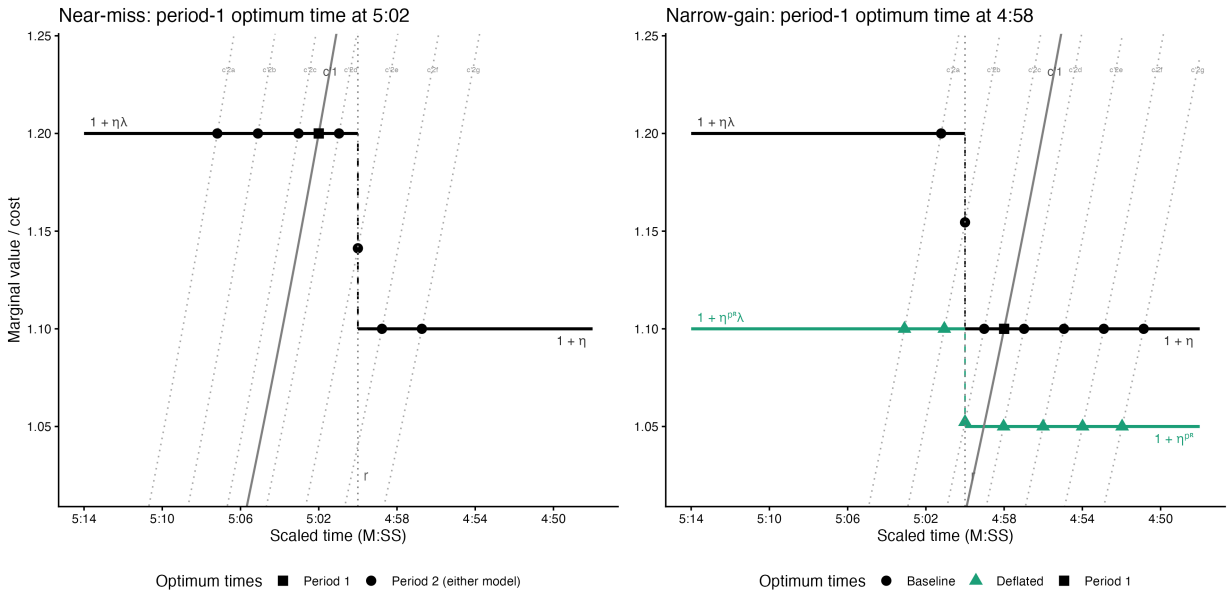


Figure A38: Illustrative equilibrium response to a family of period-2 ability shocks, linear model. Left: a runner whose period-1 equilibrium sits 2 seconds short of r (5:02, "near-miss"). Right: a runner 2 seconds past it (4:58, "narrow-gain"). Gray dotted lines are period-2 cost curves c'_{2a}–c'_{2g}, one per shock $\delta \in \{-5, -3, -1, 1, 3, 5, 7\}$ applied to the runner's underlying (unobserved) benchmark, centered on the model's true mean shock $\delta = 1$ and spanning roughly $\pm$ one SD; the solid gray line is the unshocked period-1 cost curve, c'_1. Black horizontal steps are the undeflated marginal-benefit curve (the only curve relevant in the left panel, since r is never surpassed there; the baseline curve in the right panel). The teal step in the right panel is the deflated marginal-benefit curve, which applies only once r has been surpassed. Marginal-benefit levels are labeled directly on each curve. Points mark the resulting period-2 equilibria for each shock, colored and shaped by model as in the legend; where baseline and deflated coincide (left panel), only one marker style is shown.

Table A17: Simulated and empirical moments: fixed and fitted parameters

		Baseline				Deflated			
		Fixed		Fitted		Fixed		Fitted	
Moment	Empirical	Est.	Match	Est.	Match	Est.	Match	Est.	Match
A. Bunching (Chetty statistic)									
Bunching in times	0.112	0.362	✓ (over)	0.003	×	0.260	✓ (over)	0.079	✓
Bunching in PRs	0.174	0.362	✓ (over)	0.003	×	0.420	✓ (over)	0.153	✓
PR bunching > time bunching (ratio)	∼ 1.5	0.999	×	0.892	×	1.614	✓	1.931	✓
B. RD: improvement probability									
Following times	−0.031	0.005	×	0.001	×	−0.060	✓	−0.040	✓
Following PRs	−0.040	−0.006	×	0.001	×	−0.084	✓	−0.052	✓
Greater for PRs (ratio)	∼ 1.3	−1.276	×	0.871	×	1.412	✓	1.323	✓
C. RD: time improvement, seconds									
Following times	−0.24	1.143	×	0.046	×	0.004	×	−0.304	✓
Following PRs	−0.36	1.066	×	0.049	×	−0.205	✓	−0.433	✓
Greater for PRs (ratio)	∼ 1.5	0.932	×	1.073	×	— [a]	×	1.424	✓

Notes: Fixed columns use $\lambda = 2$, $\eta = 0.1$, $\eta^{PR} = 0.035$, $\beta = 5$ (calibrated from outside literature, not fit to these moments). Fitted columns choose (η, λ) for Baseline and $(\eta, \lambda, \eta^{PR}/\eta)$ for Deflated jointly via a grid-search-initialized Nelder-Mead minimizing a weighted sum of squared deviations between simulated and empirical moments (weights: $1/(\text{PR-value})^2$ for each times/PR pair, so an equal proportional miss counts equally across moments, and an equal absolute miss counts equally between a pair's times and PR component). Fitted parameters: Baseline $\eta = 0.010$, $\lambda = 1.080$; Deflated $\eta = 0.054$, $\lambda = 1.549$, $\eta^{PR}/\eta = 0.347$. All estimates are means across 100 replicates of 50,000 runners. Bunching statistics replicate the empirical procedure exactly (quartic polynomial counterfactual over a ±25 s window with a mass-preserving level shift, normalized by the local counterfactual mass; see Appendix data notes). RD estimates use OLS with a linear spline and a ±10 s bandwidth; a negative estimate indicates that runners just surpassing the reference point (fast side) improve less in the subsequent period. For comparison rows (3, 6, 9), the estimate is the PR-based divided by the time-based statistic. *Match* reflects whether the model reproduces the direction and approximate magnitude of the empirical finding; "(over)" flags a same-signed estimate that substantially overshoots.

[a] Omitted: the underlying time-improvement estimate is indistinguishable from zero at these parameters, making the ratio numerically unstable a

A7.2.2 Simulated method of moments

The Fitted columns of Table A17 choose $(\eta, \lambda, \eta^{PR}/\eta)$ jointly, minimizing a weighted sum of squared deviations between the six primary simulated and empirical moments in Panels A–C: the bunching and regression discontinuities in improvement time and improvement probabilities, for both finishing times and PRs. This works via a grid-search-initialized Nelder-Mead search, evaluating each

candidate parameter vector across 100 replicates of 50,000 runners. Moments are weighted such that proportional differences relative to the empirical values are equally weighted.[19] The Baseline specification cannot reproduce the sign of either RD discontinuity at any parameter values, as discussed in Section A7.2.1: fitting only pushes η toward its lower bound, since any positive η pushes the time-improvement estimate the wrong way regardless of its size. The Deflated specification, by contrast, fits closely on every moment once $(\eta, \lambda, \eta^{PR}/\eta)$ are chosen jointly, including all three PR-to-times ratios. Notably, the fitted deflation ratio is $\eta^{PR}/\eta = 0.347$, well below the nested Baseline specification at $\eta^{PR}/\eta = 1$: the search thus selects a large reduction in the weight on reference dependence once the threshold has been surpassed, rather than the Baseline model. Deflation also generates the systematic gap between the PR and times moments, producing larger bunching and larger discontinuities following PRs than following times, which the Baseline model is unable to produce. In fact, the Baseline model fit drives reference dependence toward zero, returning $\eta = 0.010$ and $\lambda = 1.080$—almost no weight on the reference point and almost no loss aversion. In switching reference dependence off, the fitted Baseline model loses the bunching it was introduced to explain, inconsistent both with previous work and with the empirical data. Consistent with all of the above, both summary measures of fit in Table A18 favor the Deflated specification substantially: a weighted objective of 0.264 against 4.973, and a mean absolute percentage error across the six primary moments of 24.5% against 105.7%.

[19] Moments are weighted by $1/(\text{PR-value})^2$ within each times/PR pair, so that an equal proportional miss counts equally across moments while an equal absolute miss counts equally within a time/PR pair.

Table A18: SMM fitted parameters and overall model fit

	Baseline	Deflated
η	0.010	0.054
λ	1.080	1.549
η^{PR}/η	—	0.347
Objective (weighted SSE)	4.973	0.264
Mean abs. % error	105.7%	24.5%

Notes: Objective is the weighted sum of squared deviations between simulated and empirical moments minimized by the SMM fit (weights as in Table A17's notes: $1/(\text{PR-value})^2$ shared across each times/PR pair). Mean absolute % error averages $|(\text{simulated} - \text{empirical})/\text{empirical}|$ across the six primary moments (bunching and RD levels; excludes the three comparison ratios). Both metrics show Deflated fitting far more closely than Baseline, whose poor fit reflects a structural limitation rather than a search failure: no value of η lets the undeflated model reproduce the sign of either RD discontinuity (see below).

A7.2.3 Parameter sensitivity

Figures A39–A41 show how the bunching and regression-discontinuity estimates respond to λ, η, and the deflation ratio η^{PR}/η, each varied one at a time around the baseline values, using $S = 100$ replicates of $N = 50{,}000$ runners per parameter value. Bunching, and, for the deflated models, the improvement-probability discontinuity, move monotonically in the expected direction with all three parameters. The time-improvement discontinuity is less stable, crossing zero depending on parameter values.

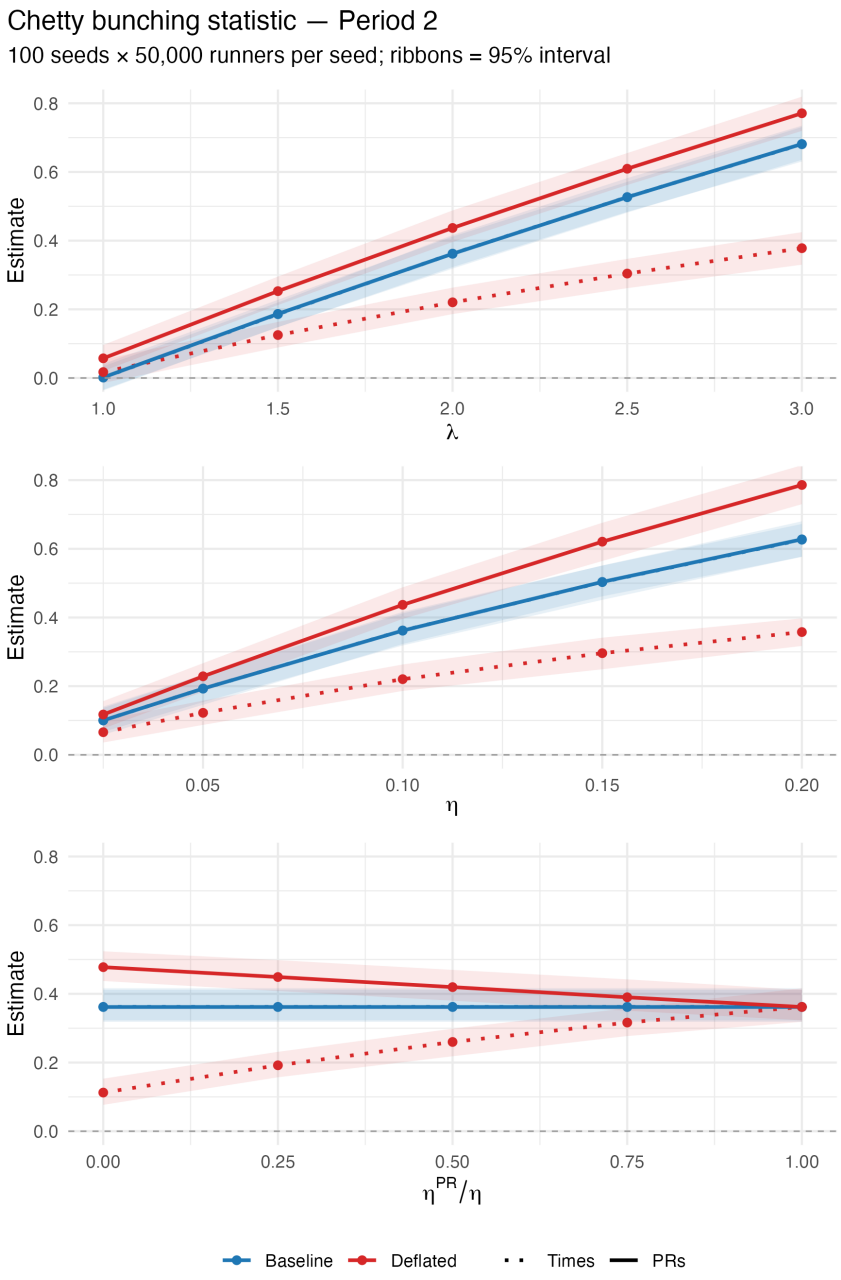


Figure A39: Sensitivity of the Chetty bunching statistic (period 2) to λ, η, and the deflation factor η^{PR}/η (top to bottom), each varied one at a time around the baseline calibration. Blue lines are the baseline model; red lines the deflated model. Dotted lines show times; solid lines show personal records (PRs) – both are overlaid on the same panel rather than shown separately. Ribbons span the 2.5th–97.5th percentile across 100 simulation replicates of 50,000 runners each.

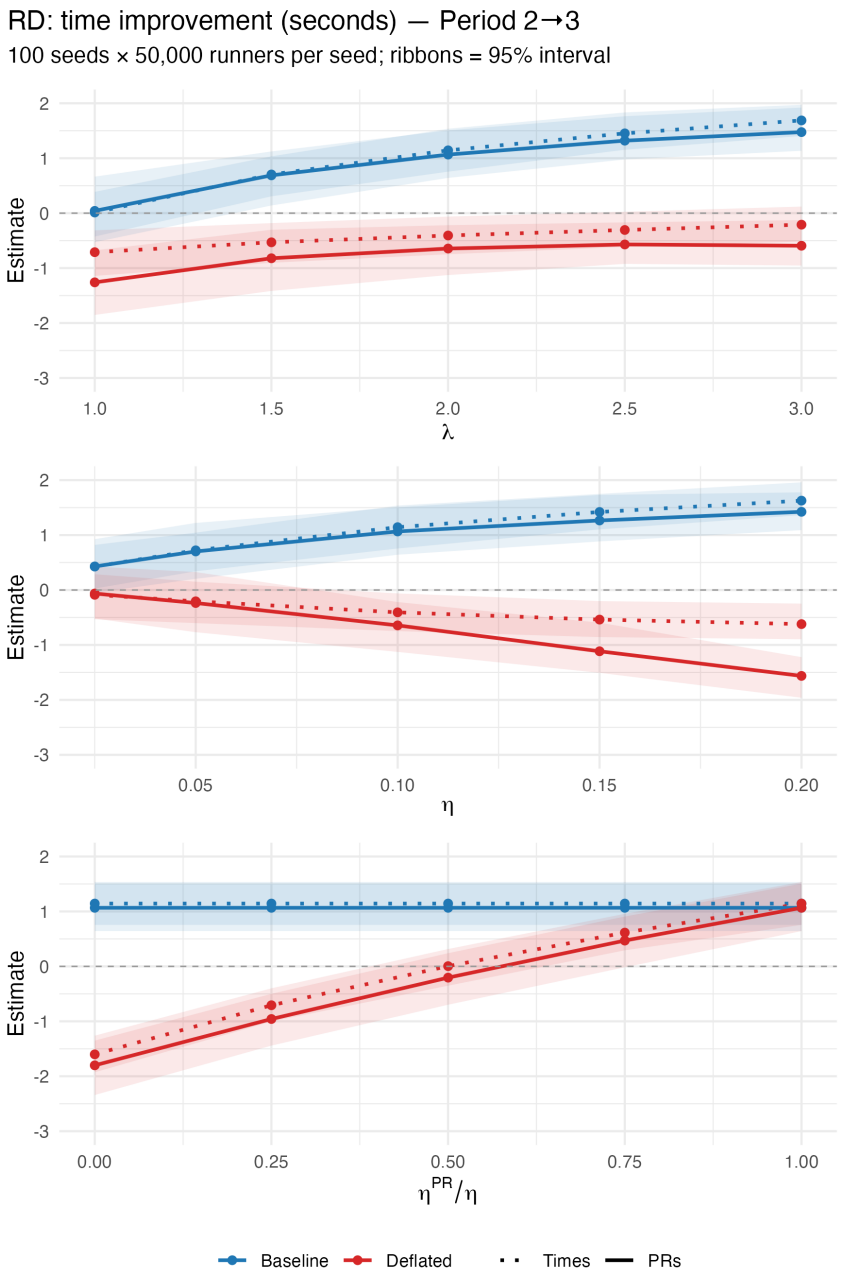


Figure A40: Sensitivity of the regression-discontinuity estimate for time improvement (seconds, period 2→3) to λ, η, and the deflation factor. A negative estimate indicates that runners on the fast side of the reference point improve less in the following period. Panels, colors, and linetypes as in Figure A39.

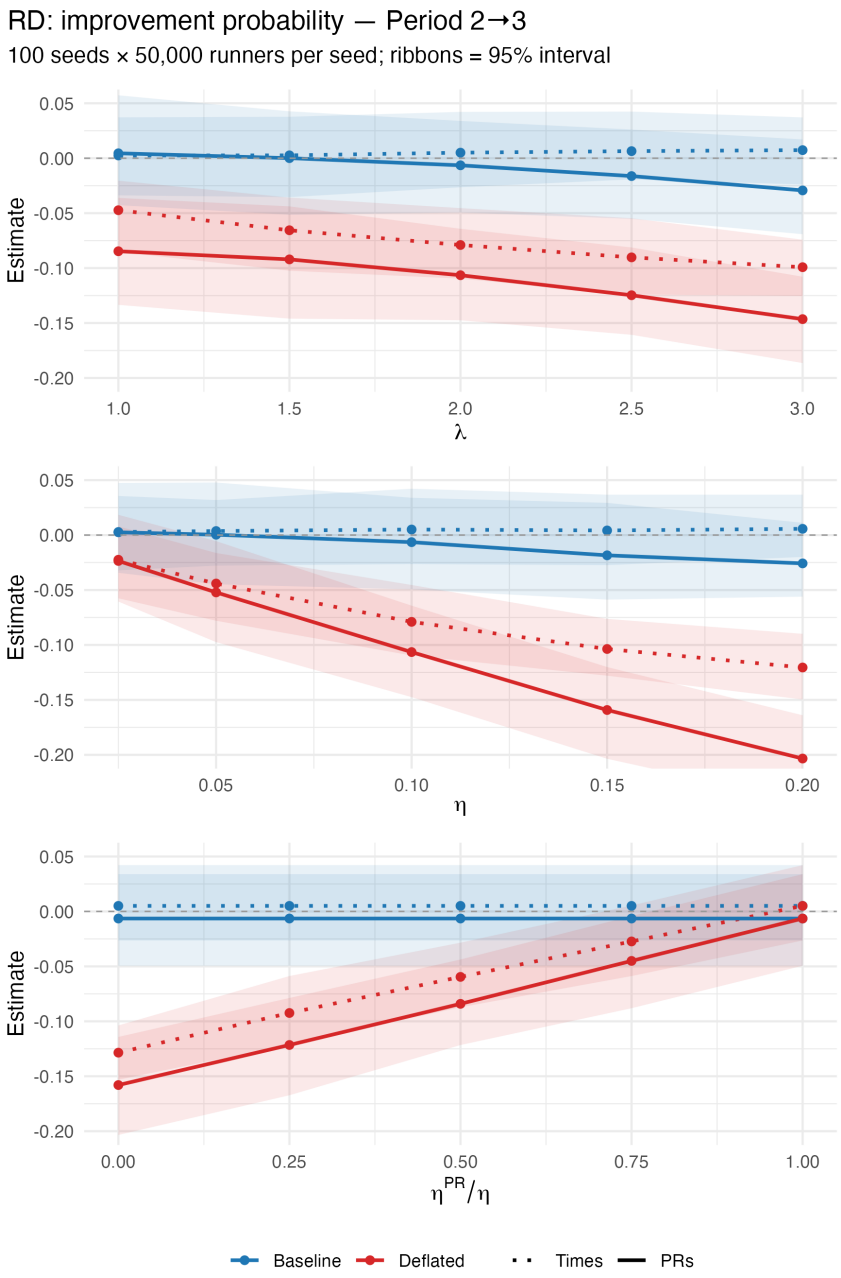
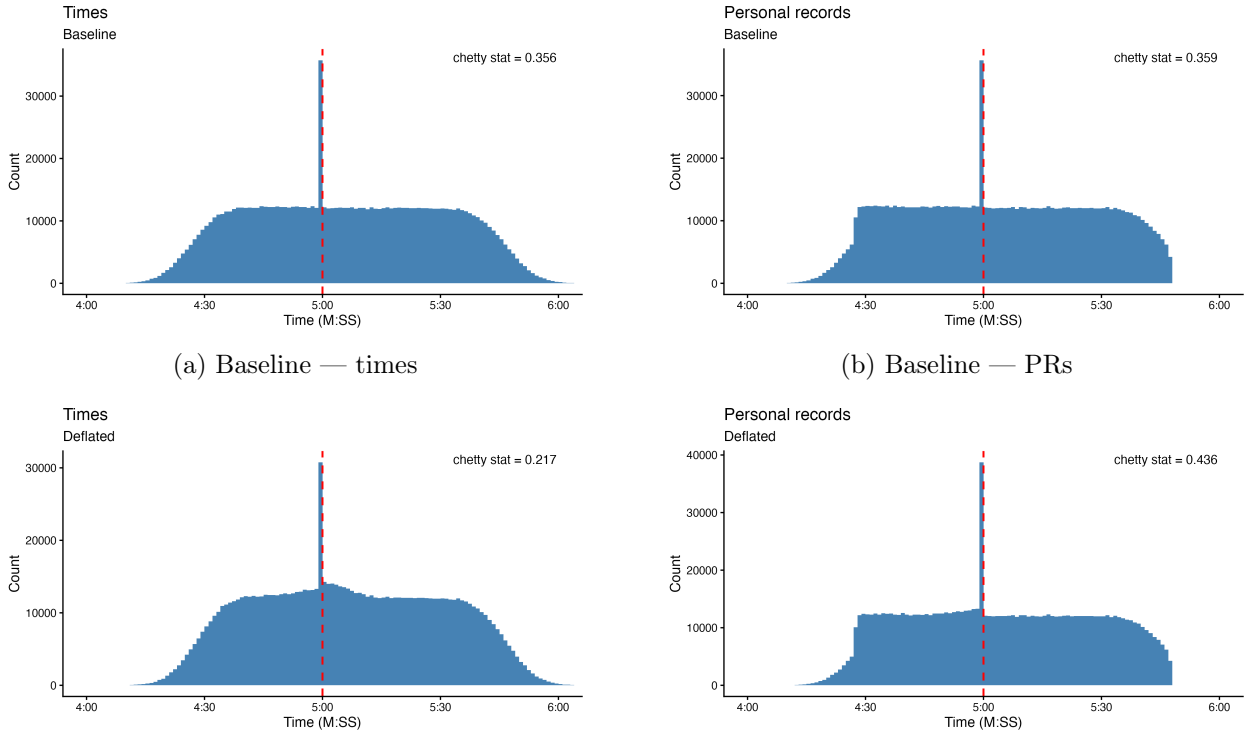


Figure A41: Sensitivity of the regression-discontinuity estimate for the probability of improvement (period 2→3) to λ, η, and the deflation factor. A negative estimate indicates that runners on the fast side of the reference point are less likely to improve in the following period. Panels, colors, and linetypes as in Figure A39.

A7.3 Additional analyses

A7.3.1 Example simulation plots

Figure A42: Simulated time and PR distributions



(a) Baseline — times

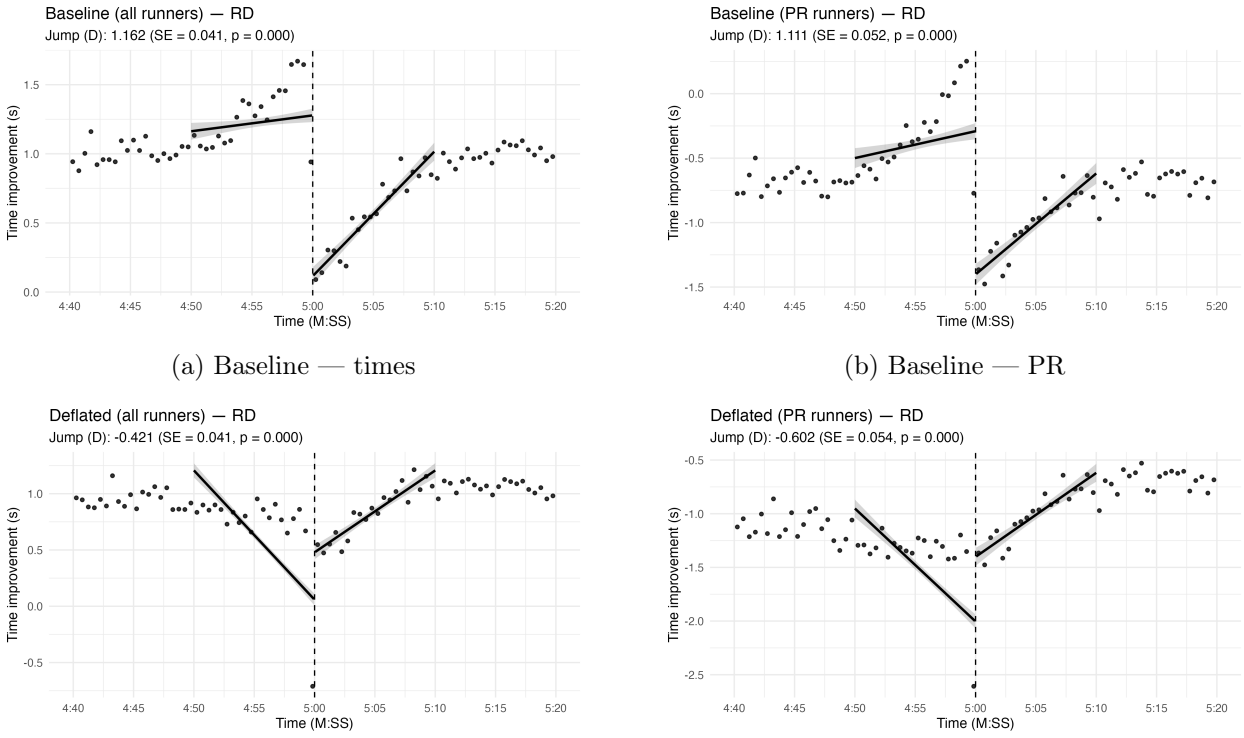
(b) Baseline — PRs

(c) Deflated — times

(d) Deflated — PRs

Notes: Distributions of period-2 times (left) and personal records (right), at each model's SMM-fitted parameters (Table A17, Fitted columns), not the fixed calibration. Bins are 1 second wide; the vertical dashed line marks the 5:00 reference point (300 s). The Chetty statistic in the top-right corner of each panel estimates normalized excess mass in the 5 s immediately above the reference point, computed identically to the estimates reported in Table A17. *Single large-sample draw;* $N = 1{,}000{,}000$ *runners.*

Figure A43: Simulated time improvement discontinuity



(a) Baseline — times

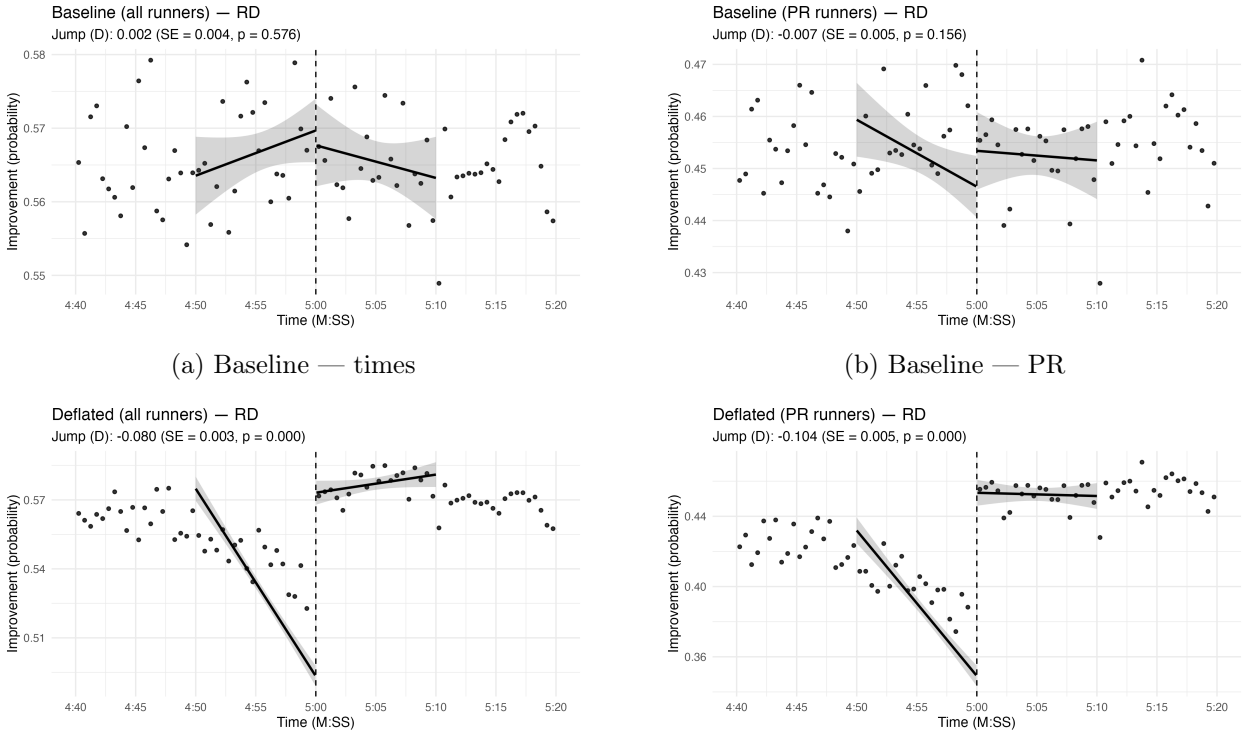
(b) Baseline — PR

(c) Deflated — times

(d) Deflated — PR

Notes: Regression-discontinuity plots for period-2-to-3 time improvement (seconds) at the 5:00 reference point, at each model's SMM-fitted parameters (Table A17, Fitted columns), not the fixed calibration. Left panels use all period-2 runners with the period-2 time as the running variable (RV); right panels restrict to new-PR runners using the PR as the running variable. Dots are binned means; lines and shaded ribbons show the OLS linear-spline fit and 95% confidence interval within a ±10 s bandwidth. The jump in each subtitle measures the difference in period-3 time improvement between runners just faster and just slower than 5:00 (positive = runners surpassing 5-min improve more), computed identically to the estimates reported in Table A17. *Single large-sample draw;* $N = 1{,}000{,}000$ *runners*

Figure A44: Simulated improvement probability discontinuity



(a) Baseline — times

(b) Baseline — PR

(c) Deflated — times

(d) Deflated — PR

Notes: Regression-discontinuity plots for the probability of improvement (period 2→3) at the 5:00 reference point, at each model's SMM-fitted parameters (Table A17, Fitted columns), not the fixed calibration. Layout identical to Figure A20. The jump measures the difference in improvement probability between runners just faster and just slower than 5:00 (negative = faster runners are less likely to improve), computed identically to the estimates reported in Table A17. *Single large-sample draw;* $N = 1{,}000{,}000$ *runners.*

A7.3.2 Robustness: alternative value functions

We consider two alternative value functions as robustness checks, neither of which is part of the main analysis: a diminishing-sensitivity specification and a probabilistic marginal benefit. Both add at least one free parameter relative to the linear model. Fixed columns below use the same outside-calibration parameters as Table A17, computed with the identical, corrected bunching/RD estimator. Fitted columns choose all free parameters jointly via the same grid-search-initialized Nelder-Mead procedure used in Table A17, with grid and simplex scaling for the diminishing-sensitivity model's extra curvature parameter α extended with the same care. Among Deflated specifications – the economically preferred ones, since only they reproduce the empirical RD signs – neither alternative improves on the linear model's fit (weighted SMM objective: linear 0.264, diminishing-sensitivity 0.360, probabilistic 0.864); diminishing sensitivity comes closest only by fitting $\alpha = 0.906$, close to the linear-nesting value $\alpha = 1$. Among Baseline specifications, both alternatives post a lower objective than linear (linear 4.973, diminishing-sensitivity 3.052, probabilistic 3.772), but this reflects added curve-fitting flexibility rather than resolved economics: like the linear model, neither alternative can reproduce the sign of the improvement-probability RD discontinuity at any parameter values.

Diminishing sensitivity A diminishing-sensitivity value function is a common specification in the literature (Tversky and Kahneman, 1992; Allen et al., 2017), in which marginal benefit is largest at the reference point and decays with distance from it on both sides, rather than jumping between two flat levels as in our linear specification (which it nests as the special case $\alpha = 1$). Table A19 uses $\alpha = 0.88$ (Tversky and Kahneman's own estimate of this curvature).

Value function. Writing $x = t - r$ for the runner's model-time distance above (gain) or below (loss) the reference point, the Tversky-Kahneman power value function is

$$v(x) = \begin{cases} x^{\alpha} & x \geq 0 \quad \text{(gain)} \\ -\lambda(-x)^{\alpha} & x < 0 \quad \text{(loss)} \end{cases}$$

and total value is $V_{ip}(t) = t + \eta_{ip}\, v(t - r)$, the same linear-in-$t$ base plus reference-dependent term

as the main specification, with the kinked-linear v replaced by this concave-in-gains, convex-in-losses power form.

Marginal value. Differentiating gives the marginal benefit used in the equilibrium condition:

$$V'_{ip}(t) = \begin{cases} 1 + \eta_{ip}\,\alpha\,(t-r)^{\alpha-1} & t > r \quad \text{(gain domain)} \\ 1 + \eta_{ip}\,\lambda\,\alpha\,(r-t)^{\alpha-1} & t \le r \quad \text{(loss domain)} \end{cases}$$

At $\alpha = 1$ both branches collapse to the linear model's $1 + \eta_{ip}$ and $1 + \eta_{ip}\lambda$ exactly, confirming the nesting. For $\alpha < 1$, V'_{ip} diverges as $t \to r$ from either side (steep sensitivity right at the reference point) and flattens toward 1 far from it (diminishing sensitivity), rather than jumping discretely as in the linear case.

Table A19: Robustness: diminishing-sensitivity model

		Baseline				Deflated			
		Fixed		Fitted		Fixed		Fitted	
Moment	Empirical	Est.	Match	Est.	Match	Est.	Match	Est.	Match
A. Bunching (Chetty statistic)									
Bunching in times	0.112	0.689	✓ (over)	0.184	✓ (over)	0.472	✓ (over)	0.115	✓
Bunching in PRs	0.174	0.686	✓ (over)	0.185	✓	0.743	✓ (over)	0.161	✓
PR bunching > time bunching (ratio)	∼ 1.5	0.996	×	1.003	×	1.574	✓	1.402	✓
B. RD: improvement probability									
Following times	−0.031	0.004	×	0.001	×	−0.079	✓	−0.024	✓
Following PRs	−0.040	0.005	×	0.002	×	−0.069	✓	−0.034	✓
Greater for PRs (ratio)	∼ 1.3	1.350	✓	1.845	×	0.884	×	1.445	✓
C. RD: time improvement, seconds									
Following times	−0.24	0.848	×	−0.058	✓	−0.450	✓	−0.113	✓
Following PRs	−0.36	0.875	×	−0.026	×	−0.206	✓	−0.209	✓
Greater for PRs (ratio)	∼ 1.5	1.032	×	0.456	×	0.457	×	1.849	✓

Notes: Diminishing-sensitivity value function; Fixed columns use $\alpha = 0.88$ (Tversky and Kahneman's estimate), $\lambda = 2$, $\eta = 0.1$, $\eta^{PR} = 0.035$, $\beta = 5$, as in Table A17, with estimates as means across 100 replicates of 50,000 runners using the identical bunching/RD estimator. Fitted columns choose (η, λ, α) for Baseline and $(\eta, \lambda, \eta^{PR}/\eta, \alpha)$ for Deflated jointly, via the same grid-search-initialized Nelder-Mead and weighting as Table A17, with the grid and simplex scaling extended to cover α. Fitted parameters: Baseline $\eta = 0.054$, $\lambda = 1.040$, $\alpha = 0.370$; Deflated $\eta = 0.043$, $\lambda = 1.613$, $\eta^{PR}/\eta = 0.333$, $\alpha = 0.906$ — close to $\alpha = 1$, i.e. close to nesting the linear specification, once deflation is allowed to do the main work. *Match* reflects whether the model reproduces the direction and approximate magnitude of the empirical finding, as in Table A17.

Probabilistic marginal benefit The linear model assumes runners can choose their finishing time precisely. In practice, however, realized performance is likely subject to uncertainty: even if a runner targets a particular time, fatigue, pacing errors, and race-day conditions introduce noise around that target. To capture this, we consider an alternative specification in which each runner chooses a *target* time $\tilde{t}_{ip}$, but the realized time is:

$$t_{ip} = \tilde{t}_{ip} + \varepsilon_{ip}, \qquad \varepsilon_{ip} \sim \mathcal{N}(0,\, \sigma_t^2),$$

with $\sigma_t = 1$ second. The runner now maximizes expected value, $\mathbb{E}[V(t_{ip})]$, by choosing $\tilde{t}_{ip}$.

Expected marginal value. Because ε_{ip} is normally distributed, the probability that the realized time falls in the loss domain can be derived from the standard normal CDF. Taking the first-order condition with respect to $\tilde{t}_{ip}$ yields a smooth, probability-weighted average of marginal benefits of both gains and losses.

$$\mathbb{E}\big[V'_{ip}(\tilde{t}_{ip})\big] = (1 + \eta_{ip}) + \eta_{ip}(\lambda - 1)\,\Phi\!\left(\frac{r - \tilde{t}_{ip}}{\sigma_t}\right).$$

Given relatively limited noise, this function is similar to the linear model for most of the support, but replaces the discrete jump of the linear specification with a smooth S-curve in the region around r.

Equilibrium. The runner's target time solves:

$$A_{ip}\,\tilde{t}_{ip}^{\,\beta} = (1 + \eta_{ip}) + \eta_{ip}(\lambda - 1)\,\Phi\!\left(\frac{r - \tilde{t}_{ip}}{\sigma_t}\right).$$

Table A20: Robustness: probabilistic marginal benefit

		Baseline				Deflated			
		Fixed		Fitted		Fixed		Fitted	
Moment	Empirical	Est.	Match	Est.	Match	Est.	Match	Est.	Match
A. Bunching (Chetty statistic)									
Bunching in times	0.112	0.094	✓	0.002	×	0.066	✓	0.046	✓
Bunching in PRs	0.174	0.114	✓	0.002	×	0.160	✓	0.123	✓
PR bunching > time bunching (ratio)	∼ 1.5	1.219	✓	0.914	×	2.442	✓ (over)	2.698	✓
B. RD: improvement probability									
Following times	−0.031	−0.019	✓	−0.002	×	−0.061	✓ (over)	−0.046	✓
Following PRs	−0.040	−0.014	×	−0.005	×	−0.089	✓ (over)	−0.064	✓
Greater for PRs (ratio)	∼ 1.3	0.753	×	2.339	✓	1.464	✓	1.391	✓
C. RD: time improvement, seconds									
Following times	−0.24	0.520	×	−0.034	×	−0.317	✓	−0.221	✓
Following PRs	−0.36	0.571	×	−0.048	×	−0.728	✓ (over)	−0.480	✓
Greater for PRs (ratio)	∼ 1.5	1.098	×	1.426	✓	2.298	✓ (over)	2.173	✓

Notes: Probabilistic marginal benefit with $\sigma_t = 1$ (held fixed, not fit, matching how β is held fixed in the main model). Fixed columns use $\lambda = 2$, $\eta = 0.1$, $\eta^{PR} = 0.035$, $\beta = 5$, as in Table A17, with estimates as means across 100 replicates of 50,000 runners using the identical bunching/RD estimator. Fitted columns choose (η, λ) for Baseline and $(\eta, \lambda, \eta^{PR}/\eta)$ for Deflated jointly, via the same grid-search-initialized Nelder-Mead and weighting as Table A17. Fitted parameters: Baseline $\eta = 0.385$, $\lambda = 1.001$; Deflated $\eta = 0.051$, $\lambda = 2.554$, $\eta^{PR}/\eta = 0.418$. As in the linear model, Baseline cannot reproduce the sign of the improvement-probability discontinuity at any parameter values; unlike the linear model, fitting here does not push η toward its lower bound, since the smooth probabilistic kink lets a larger η partly compensate on the bunching and time-improvement moments even though it cannot fix the improvement-probability sign. *Match* reflects whether the model reproduces the direction and approximate magnitude of the empirical finding, as in Table A17.