Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Reference-Dependent Preferences: Evidence from Marathon Runners

Eric J. Allen, Patricia M. Dechow, Devin G. Pope, George Wu

To cite this article:
Eric J. Allen, Patricia M. Dechow, Devin G. Pope, George Wu (2017) Reference-Dependent Preferences: Evidence from Marathon Runners. Management Science 63(6):1657-1672. https://doi.org/10.1287/mnsc.2015.2417

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 63, No. 6, June 2017, pp. 1657–1672
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Reference-Dependent Preferences: Evidence from Marathon Runners

Eric J. Allen,[a] Patricia M. Dechow,[b] Devin G. Pope,[c] George Wu[c]

[a] Marshall School of Business, University of Southern California, Los Angeles, California 90089; [b] Haas School of Business, University of California, Berkeley, Berkeley, California 94705; [c] Booth School of Business, University of Chicago, Chicago, Illinois 60637

Contact: eric.allen@marshall.usc.edu (EJA); patricia_dechow@haas.berkeley.edu (PMD); dpope@chicagobooth.edu (DGP); wu@chicagobooth.edu (GW)

Received: March 30, 2015
Revised: August 17, 2015
Accepted: October 27, 2015
Published Online in Articles in Advance: April 20, 2016

https://doi.org/10.1287/mnsc.2015.2417

Abstract. Theories of reference-dependent preferences propose that individuals evaluate outcomes as gains or losses relative to a neutral reference point. We test for reference dependence in a large data set of marathon finishing times (n = 9,789,093). Models of reference-dependent preferences such as prospect theory predict bunching of finishing times at reference points. We provide visual and statistical evidence that round numbers (e.g., a four-hour marathon) serve as reference points in this environment and as a result produce significant bunching of performance at these round numbers. Bunching is driven by planning and adjustments in effort provision near the finish line and cannot be explained by explicit rewards (e.g., qualifying for the Boston Marathon), peer effects, or institutional features (e.g., pacesetters).

History: Accepted by John List, behavioral economics.
Funding: The authors thank the John Templeton Foundation (New Paths to Purpose project) for generous financial support.
Supplemental Material: Data and the online appendix are available at https://doi.org/10.1287/mnsc.2015.2417.

Keywords: reference dependence • prospect theory • loss aversion • bunching • effort provision

1. Introduction

Recent theories of economic behavior propose that the evaluation of an outcome may be affected by comparisons of that outcome with a reference point and not merely tastes, risk attitudes, and wealth levels, as in classical economic models. For example, how an employee views a bonus of $1,000 might depend on the level of previous bonuses, what bonuses were distributed to other members of the organization, or the employee's expectations about what bonuses were possible (Card et al. 2012, Kahneman 1992, Kőszegi and Rabin 2006).

A reference point divides outcomes into gains or losses, thus creating a qualitative difference in the valuation of outcomes slightly above or below that reference point. We suggest that the distinguishing feature of reference-dependent models is some form of discontinuity at the reference point that is psychologically based and not the result of an extrinsic benefit. For example, a primary feature of prospect theory, the most well-known and influential account of reference-dependent preferences, is loss aversion (Kahneman and Tversky 1979, Tversky and Kahneman 1992). The premise that "losses loom larger than gains" (Kahneman and Tversky 1979) has implications for a wide range of economic activities, including risky decision making, choice of consumption bundles, and effort provision (DellaVigna 2009, Tversky and Kahneman 1991). A second property, diminishing sensitivity, is captured by prospect theory's characteristic S-shaped value function that is concave for gains and convex for losses. Although prospect theory is the most prominent model of reference dependence, the discontinuity at the reference point in some instances might instead be produced by a jump (or "notch") in the utility function at the reference point.

Researchers have moved beyond Kahneman and Tversky's laboratory demonstrations of reference dependence to explain behavioral anomalies across a wide variety of field settings.[1] In a recent review of prospect theory, Barberis (2013) highlighted the key challenge to researchers testing for field evidence of reference-dependent preferences: it is often difficult to know exactly what reference points are relevant for individuals in field settings. The difficulty in identifying the appropriate reference point is best illustrated by a stream of work examining the possible role that reference points play in labor supply and effort provision. Camerer et al. (1997) argued that taxi drivers have a downward-sloping labor supply curve induced by daily income targets (see also Fehr and Goette 2007, and Mas 2006). This paper led to additional analyses that used different data sets and econo-

metric methods to examine if taxi drivers indeed have reference-dependent preferences, with some arguing against (Farber 2005, 2008, 2015) and some arguing in favor (Ashenfelter et al. 2010, Crawford and Meng 2011). The primary empirical challenge in these papers has been modeling reference points that are unobservable, heterogeneous, and possibly nonstationary.

In this paper, we test for reference dependence in a data set of almost 10 million marathon finishing times. For several reasons, marathon running is an ideal environment to look for field evidence of reference dependence. First and most importantly, we propose that there are clear and stable reference points in this setting. Specifically, we cite survey evidence that the majority of runners judge their performance relative to round numbers (e.g., running a marathon in four hours). The prevalence across runners of round number reference points provides us with a cleaner and sharper test of reference dependence than other settings where reference points are unknown or likely to differ across individuals (e.g., taxi drivers). Coupled with our large sample, these universal reference points allow us to very easily and credibly identify evidence of reference dependence using nonparametric methods. The richness of our data also allows us to examine how reference points impact effort provision at different points in the race.[2]

Second, these round number reference points are largely not tied to external rewards and thus, we argue, primarily provide psychological, not economic, motivation. That said, runners may also care about how their finishing times are perceived by others. If this is the case, evidence for reference dependence may not reveal an intrinsic reference point but instead reflect an audience effect in which runners feel that they will be evaluated more favorably if they run just faster than a round number. For example, a runner may feel significantly better about herself if she runs a 3:59 marathon rather than a 4:01 marathon (an intrinsic reference point), or she may feel that other people will be demonstrably more impressed with a 3:59 than a 4:01 marathon (an audience effect). We make two comments about this critique. First, most studies of reference dependence are unable to distinguish between intrinsic versus audience effects. For example, does a taxi driver care about reaching a particular target, or is she worried about her spouse's reaction? Is a homeowner reluctant to sell his home for less than his purchase price because he will feel a loss or because of concern about what neighbors will think if they discover that he lost money on this transaction? Like these other studies, we acknowledge that part of the effect may be driven by others who evaluate outcomes relative to round numbers.[3] Second, and most important, all of these examples still reflect reference-dependent evaluations, regardless of whether the reference dependence originated with the runner, taxi driver, or homeowner or that agent's audience.[4]

Consistent with a simple model of reference-dependent preferences and in sharp contrast with the predictions from a standard model of utility, we find a lumpy distribution of finishing times, with bunching just ahead of round numbers. For example, 50.0% more runners finished in the minute just under three hours than the minute just over three hours. We observe qualitatively similar patterns for all relevant 60-minute marks as well as 30-minute marks and many 10- and 15-minute marks. We provide evidence that this effect is primarily psychological and cannot be explained by financial incentives or other extrinsic rewards (e.g., qualifying for the Boston Marathon) or by institutional features (e.g., pacesetters) and show that this effect is explained in part by pacing and planning and in part by effort provision over the final stages of the marathon. Runners are more likely to speed up and less likely to slow down when they are on pace to finish just ahead of a round number reference point.

Our paper proceeds as follows. In Section 2, we present a simple model that demonstrates how reference-dependent preferences such as those defined in prospect theory will produce bunching in running performance at reference points in a similar way to how taxpayers bunch at a kink in the tax code. In Section 3, we discuss some institutional features of marathons and describe our data. We present the main results in Section 4. We conclude the paper in Section 5 with a brief discussion of the broader significance of our findings.

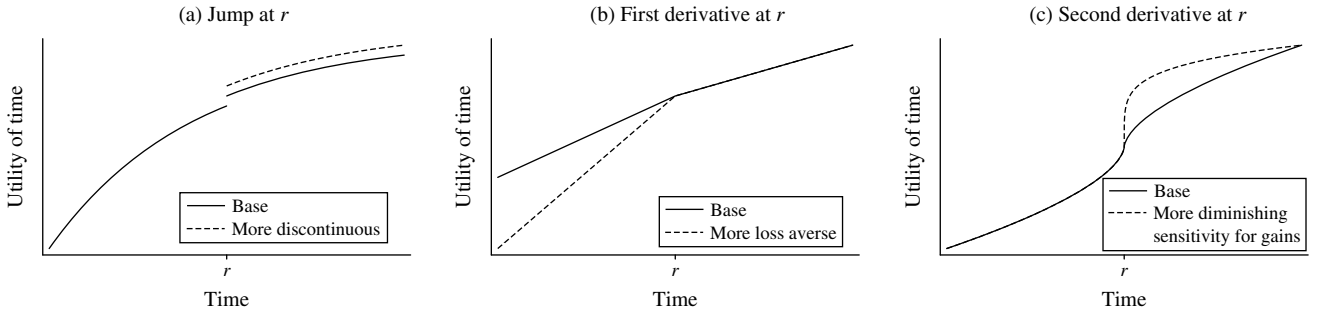
2. Conceptual Framework

We show how a simple model of reference dependence produces bunching of performance at a reference point. Reference dependence implies that individuals evaluate outcomes just above or just below the neutral reference point in a manner that is inconsistent with standard utility theory. The qualitatively distinct perception of outcomes that are just above or below a reference point can take several forms. Let $v_r(\cdot)$ denote a utility function that shows reference dependence around a reference point r. Below, we detail three primary forms of reference dependence:

1. A jump or discontinuity at r: $\lim_{\epsilon\to 0} v_r(r+\epsilon) \neq \lim_{\epsilon\to 0} v_r(r-\epsilon)$.
2. A kink or discontinuity in the first derivative at r: $\lim_{\epsilon\to 0} v_r'(r+\epsilon) \neq \lim_{\epsilon\to 0} v_r'(r-\epsilon)$.
3. A kink or discontinuity in the second derivative at r: $\lim_{\epsilon\to 0} v_r''(r+\epsilon) \neq \lim_{\epsilon\to 0} v_r''(r-\epsilon)$.

The first form of reference dependence leads to a discontinuity or jump in the utility function. In tax settings, this has been called a "notch" (Kleven and Waseem 2013). Such a jump is featured in models of level of aspiration that have appeared in both psychology (March and Shapira 1987) and economics

Figure 1. Three Forms of Reference-Dependent Preferences



(Diecidue and Van De Ven 2008, Fishburn 1977). For example, Diecidue and Van De Ven (2008) axiomatize an expected utility representation with a jump at the reference point, assuming that the decision maker is concerned with the probability of reaching a reference point or, in their language, aspiration level. The final two forms of reference dependence follow from prospect theory's characteristic S-shape. Loss aversion and diminishing sensitivity are special cases of the utility having a discontinuous first and second derivative (Tversky and Kahneman 1992). The solid lines in Figure 1 show discontinuities of all three forms. Markle et al. (2015) found evidence for all three forms of reference dependence in a survey of marathon runners: satisfaction as a function of performance relative to a runner's time goal exhibited loss aversion and diminishing sensitivity as in prospect theory, as well as a jump in satisfaction at the goal.

Below we show that each of these three forms of reference dependence can independently produce bunching near the reference point. The proofs for all propositions are found in Section A.2 of the online appendix. Let τ denote a runner's finishing time, and $t = k - \tau$ indicate the amount of time that a runner is faster than the worst possible finishing time k. (For expository clarity, we redefine performance so that agents are maximizing, rather than minimizing, time.) We assume that an individual has a utility function that is additively separable in benefits, $b(t)$, and costs, $c(t)$; i.e., $U(t) = b(t) - c(t)$. We assume that the benefit function is increasing in t and has at least one of the three forms of reference-dependent preferences described above. Furthermore, we assume that $c(t) > 0$ and $c'(t) > 0$; i.e., costs are positive and increasing. Throughout, we take agents to be optimizers, choosing t to maximize $U(t)$, denoting $t^*(c(t), b(t))$ to be the maximum performance for an agent with cost function $c(t)$ and benefit function $b(t)$.

One convenient way to model the heterogeneity in performance across runners is to posit a family of cost functions, $c_1(t), \ldots, c_N(t)$, where each cost function captures the abilities and preparation of each of N runners, as well as features of the marathon course, weather, etc. In contrast, we assume homogeneity in the benefit function but perform comparative statics on $b(t)$ along the three dimensions of reference dependence, looking across the family of cost functions. In each case, the comparative statics show that bunching above the reference point increases as the relevant discontinuity becomes more severe. The resulting distribution of performance will be in sharp contrast to the smooth distribution that is produced by the well-behaved cost and benefit functions assumed in standard economic models (e.g., Prendergast 1999).

We first formalize the notion of bunching by identifying, for a particular benefit function, the set of cost functions or set of individual runners in which performance exceeds the reference point by δ or less.

Definition (δ-Bunching)**.** For a particular benefit function, $b(t)$, a set of cost functions, $C(\delta, b(t))$, exhibits δ-bunching around reference point r if, for all $c(t) \in C(\delta, b(t))$, $0 \le t^*(c(t), b(t)) - r \le \delta$.

It is clear that a jump or notch at the reference point will lead to bunching exactly at the reference point. We provide a simple definition of "more discontinuous" at reference point r. This notion is depicted in panel (a) of Figure 1 as the difference between the solid and dotted curves. The more discontinuous benefit function is identical to the less discontinuous function except for a shift at the reference point.

Definition (More Discontinuous at r)**.** A benefit function $b_1(t)$ is more discontinuous at reference point r than $b_2(t)$ if $\lim_{\epsilon \to 0} b_1(r+\epsilon) - \lim_{\epsilon \to 0} b_1(r-\epsilon) > \lim_{\epsilon \to 0} b_2(r+\epsilon) - \lim_{\epsilon \to 0} b_2(r-\epsilon)$ and $b_1'(t) = b_2'(t)$ for all $t \neq r$ and $b_1(t) = b_2(t)$ for $t < r$.

Proposition 2.1. *Let $b_1(t)$ be more discontinuous at r than $b_2(t)$. Then for all $\delta > 0$, $C(\delta, b_2(t)) \subseteq C(\delta, b_1(t))$.*

Here, a psychological jump in utility plays a similar role as would monetary incentives at performance thresholds (e.g., Asch 1990, Murphy 2000, Oyer 1998).

We next turn to a discontinuity in the first derivative at r. We assume that this discontinuity reflects loss aversion. Although researchers have proposed a number of definitions of loss aversion, we use a relatively standard one: an agent is loss averse if $b'(r+\epsilon) < b' \cdot$

$(r-\epsilon)$ for all $\epsilon > 0$ (Wakker and Tversky 1993); i.e., the benefit function is everywhere steeper in losses than for the comparable gains. We first define the notion of a benefit function exhibiting more loss aversion than another benefit function.

Definition (More Loss Aversion)**.** A benefit function $b_1(t)$ is more loss averse than $b_2(t)$ if $b_1(t) = b_2(t)$ and $b_1'(-t) > b_2'(-t)$ for all $t > r$.

The definition requires that the benefit functions coincide for gains but that $b_1(t)$ be steeper than $b_2(t)$ everywhere in losses (see panel (b) of Figure 1).

The following proposition shows that this straightforward definition of "more loss averse" is related to bunching of performance above the reference point.

Proposition 2.2. *Let $b_1(t)$ and $b_2(t)$ exhibit loss aversion with $b_1(t)$ more loss averse than $b_2(t)$. Then, for all $\delta > 0$, $C(\delta, b_2(t)) \subseteq C(\delta, b_1(t))$.*

We interpret Proposition 2.2 as indicating that, as loss aversion increases, more individuals will bunch just above the reference point r. Intuitively, loss aversion increases the marginal benefit of a unit of time short of the reference point, thus boosting the motivation to get into gains.[5]

We next show that a specific discontinuity in the second derivative of the utility function, diminishing sensitivity in gains, can also lead to bunching.

Definition (More Diminishing Sensitivity in Gains). A benefit function $b_1(t)$ shows more diminishing sensitivity in gains than $b_2(t)$ on $(r, r+\delta)$ if $b_1''(t) < 0$ and $b_2''(t) < 0$ for $t > r$, $b_1(t) = f(b_2(t))$ for $t \in [r, r+\delta]$, and $b_1(t) = b_2(t)$ for $t \notin (r, r+\delta)$, where $f(\cdot)$ is a continuous strictly concave function.

See panel (c) of Figure 1 for a depiction of this property. This property requires that the benefit functions, $b_1(t)$ and $b_2(t)$, coincide except on an interval $(r, r+\delta)$. In that interval, $b_1(t)$ is strictly more concave than $b_2(t)$. Note that the proposition requires that $b_1(r+\delta) = b_2 \cdot (r+\delta)$ so that the cumulative benefits on $[r, r+\delta]$ are the same for both benefit functions.

Proposition 2.3. *Let $b_1(t)$ exhibit more diminishing sensitivity for gains than $b_2(t)$ on $(r, r+\delta)$. Then $C(\delta, b_2(t)) \subseteq C(\delta, b_1(t))$.*

The intuition of Proposition 2.3 is straightforward. More diminishing sensitivity in gains decreases the marginal benefit of running faster and therefore leads more runners to slack off once they have achieved their reference point, thus leading to bunching just above the reference point. Of course, prospect theory's characteristic S-shaped value function also involves diminishing sensitivity of losses. Our theoretical treatment has only focused on gains because more diminishing sensitivity of gains (i.e., more discontinuous at r and more loss averse) increases bunching above the reference point. However, it is straightforward to show that more diminishing sensitivity in losses results in less mass just below the reference point.

Note that our simple model does not involve any uncertainty about a runner's finishing time. Of course, a marathon runner does not know his or her actual cost function on a particular day. Although incorporating uncertainty and risk preferences into this framework will generate similar comparative statics, a model with uncertainty will have implications for the specific shape of the finishing time density function, in particular producing more diffuse bunching behavior.

The conceptual framework we have laid out in this section examines three different manifestations of reference dependence. Each form involves a discontinuity at the reference point of some kind. Our framework suggests that the distribution of marathon finishing times should be smooth if the distribution of cost functions is smooth and the benefit function is well behaved as is commonly assumed in standard economic models. We have illustrated, however, that reference-dependent preferences of any of the three forms outlined above can produce bunching or excess mass just above a reference point, even if the family of cost functions is smooth. We further proved that the amount of bunching is weakly increasing in the degree of the discontinuity at the reference point. In Section 4, we directly test for evidence of bunching at round number reference points. In Section A.9 of the online appendix, we calibrate a simple model of prospect theory and show that the observed amount of excess mass at the reference points is consistent with parameters that have previously been estimated in the literature.

3. Institutional Setting and Data

The marathon is a 42.195-kilometer (26.2-mile) road race that is popular with both professional athletes and recreational runners. Approximately 1,100 marathons were held in the United States during 2013, with an estimated 541,000 finishers (Running USA 2014). The vast majority of runners receive no financial compensation for their performance. For example, the 2013 Chicago Marathon had a prize pool of $487,000 distributed across 40 finishers over 8 divisions. The slowest prize winner finished 721st (or in the top 1.8%) out of 39,122 finishers. The race also offered time bonuses, with the slowest time bonus winner finishing 189th (or in the top 0.5% of finishers). Thus, we suggest that, for the overwhelming majority of runners, finishing times are a source of internal pride and fulfillment and not an extrinsic reward.

An important technological innovation in marathon running is a radio frequency identification (RFID) chip that is attached to a runner's shoelace or running bib and thus precisely measures a runner's finishing time.

For large marathons, many runners do not cross the start line until many minutes after the official start (e.g., it took runners in the 2013 Berlin Marathon an average of 15.59 minutes to reach the start line). The computer chip registers when a runner crosses the start line, the finish line, and various intermediate points on the course (often at 10, 20, 30, and 40 kilometers and at the half marathon). The "chip time" is the difference between when a runner reaches the start line and when a runner crosses the finish line, whereas the "clock time" is the difference between when the race starts and when a runner crosses the finish line. For most races, the chip time is regarded as the official time. Runners, therefore, usually start their watches when they cross the start line and consult their watches to check their elapsed time at various points in the race. Given that we will be testing for bunching that occurs in marathon finishing times, it is very important to have precise data. For example, self-reported data may produce bunching simply due to rounding that is common in self-reports. The available chip data are therefore essential for our purposes.[6]

The data used in this paper were obtained from various public websites.[7] In total, we have finishing times for 9,789,093 marathon finishes. The full sample contains data from 1970–2013 (88.97% of data are from 2000 or later) for 6,888 different marathon-years. For some of our analysis, we will focus on a smaller sample of 873,674 finishing times with complete 10-kilometer, half-marathon, 30-kilometer, and 40-kilometer split times. We refer to this smaller sample as the "full-split sample." The more detailed data in this smaller sample will allow us to examine some mechanisms driving the bunching of finishing times. Table 1 provides summary statistics for our full sample, as well as the full-split sample. The average finishing time is 4 hours and 26 minutes and 33 seconds (4:26:33 for short) for the full sample and 4:41:52 for the full-split sample.[8] Our marathon sample includes multiple years of all of the 50 largest U.S. marathons (as measured by 2013 rank), a relatively complete sample of all U.S. and Canadian marathons from 2000 to 2013, and several large marathons from Europe, South America, Africa, Asia, and Australia.

A runner may evoke many potential reference points for judging his or her marathon performance. For example, a runner may compare his or her finishing time to the finishing time of a close relative or friend, the average time for other people of that runner's age and or gender, the time equivalent of running 8-minute miles or 5-minute kilometers throughout the marathon, or any number of other finishing times that happen to be relevant for a particular runner.[9] We focus on round numbers (e.g., four-hour marathon time) as reference points for two primary reasons. First, these reference points are knowable to researchers, unlike, for example, the finishing time of a close friend. Second, round numbers are frequently mentioned as goals by marathon runners themselves. For example, Sackett et al. (2015) asked marathoners running 15 major U.S. marathons from 2007 to 2009 for their specific time goals. Of marathoners in that study, 86.2% indicated that they had a time goal. Of these individuals, 27.5%, 48.5%, 63.8%, and 67.2% had time goals that were divisible by 60, 30, 15, and 10 minutes, respectively. Thus, a significant fraction of marathon runners have round number time goals. In that sample, 25.5% of runners ran faster than their time goal, indicating that time goals were on average optimistic.

Below we test whether performance exhibits reference-dependent bunching around these round numbers.

4. Results

4.1. Excess Mass at Round Numbers

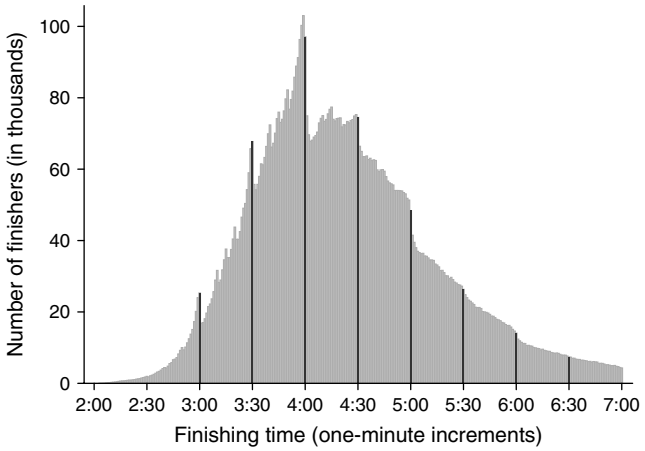
Figure 2 provides the distribution of finishing times (with 1-minute bins) for our full sample of runners.[10] The highlighted bars are the bins in the minute just before every 30-minute mark (e.g., 3:59:00–3:59:59) and indicate clear excess mass just to the left of the 30-minute marks. For example, there are 100,294, 103,018, and 97,012 finishers in the minute before the 3:58, 3:59, and

Table 1. Summary Statistics for Full Sample and Full-Split Sample

	Full sample			Full-split sample		
	Mean	Standard deviation	Observations	Mean	Standard deviation	Observations
Finishing time (HH:MM:SS)	4:26:33	0:59:11	9,789,093	4:41:52	1:04:37	873,674
Marathon year	2,005.88	6.49	9,789,093	2,009.21	2.50	873,674
Age	39.03	10.85	5,403,441	38.89	10.34	645,521
Male (1 = Male, 0 = Female)	0.66	0.48	8,061,134	0.61	0.49	812,152
Split 10 kilometers (HH:MM:SS)	1:02:22	0:17:54	2,103,830	1:01:53	0:13:14	873,674
Split half marathon (HH:MM:SS)	2:09:19	0:28:17	3,296,656	2:11:49	0:29:05	873,674
Split 30 kilometers (HH:MM:SS)	3:12:25	0:44:51	1,508,388	3:12:26	0:43:43	873,674
Split 40 kilometers (HH:MM:SS)	4:25:04	1:01:09	1,040,154	4:26:22	1:01:24	873,674

Notes. The full-split sample includes the marathons from the full sample with complete 10-, 30-, and 40-kilometer splits, as well as half-marathon splits. See http://faculty.chicagobooth.edu/george.wu/research/marathon/list.htm for a full list of marathons.

Figure 2. Distribution of Marathon Finishing Times ($n = 9{,}789{,}093$)



Note. The dark bars highlight the density in the 1-minute bin just before each 30-minute threshold.

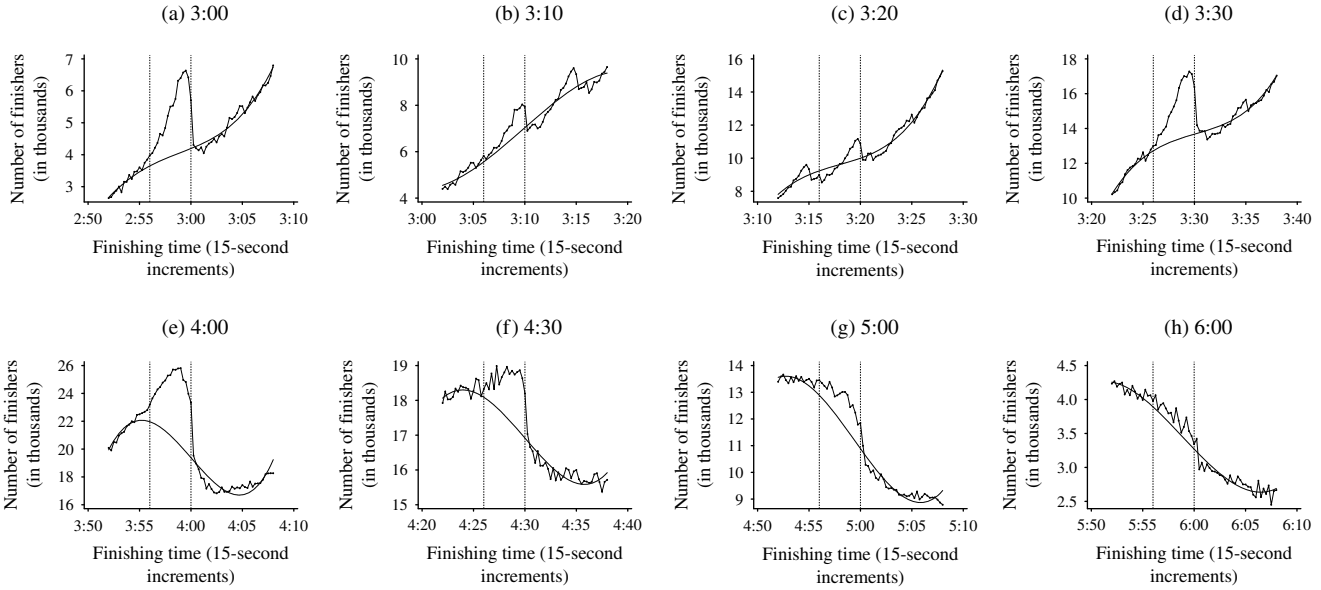
4:00 marks, compared to 74,968, 69,648, and 67,861 finishers in the 4:00, 4:01, and 4:02 bins. Although the four-hour mark is particularly dramatic, qualitatively similar differences exist at other hour and half-hour marks and, to a lesser extent, at 10- and 15-minute marks. There are 50.0%, 21.5%, and 29.5%, more finishers in the 1-minute bin before 3:00, 3:30, and 4:00, respectively, than the 1-minute bin after these round numbers. This excess mass measure for 10-minute marks is less dramatic but still substantial: 11.9%, 8.5%, 9.4%, and 7.0% for 3:10, 3:20, 3:40, and 3:50, respectively.[11]

We next measure the amount of excess mass around the round number reference point and test whether the excess mass is statistically significant by adapting a methodology proposed in Chetty et al. (2011) to quantify the extent of excess mass in an interval around a round number.[12] We draw an analogy between our setting and individual taxpayer responses to "kinks" in the tax code (e.g., Kleven and Waseem 2013, Saez 2010). Consistent with the hypothesis that income will bunch around tax rate thresholds, Chetty et al. (2011) found that Danish taxpayers bunch around the income cutoff for the top marginal income tax rate. In our setting, we hypothesize that round number reference points serve as a discontinuity in a marathoner's utility function in a similar manner to how income thresholds do for taxpayers (see Section 2). As in Chetty et al. (2011), the observed bunching is likely to be diffuse rather than a point mass. Runners are unable to perfectly control their effort levels over the course of the race. They may underestimate the amount of energy they have left, incorrectly calculate the required pace to meet the benchmark, or build a cushion into their pacing that causes them to beat the reference point by more than a small amount. As a result, rather than seeing a sharp increase in runners just beating the reference point and then an immediate drop (as required by Propositions 2.1 and 2.2), we expect to see somewhat diffuse bunching of finishing times around the reference point. This dispersion will reflect runners who attempt to meet the reference point and just miss, as well as those who beat it by a few minutes.

To calculate the amount of bunching, we follow the Chetty et al. (2011) methodology. The counterfactual distribution is estimated by fitting a quintic polynomial to the local density of finishing times around the reference point excluding the bunching region. The difference between the actual density in the bunching region and the fitted counterfactual density is the excess number of finishers around the reference point, with the standard error for the amount of excess mass determined by a bootstrap procedure. Throughout, we take the local window around each potential round-number reference point to be 16 minutes (8 minutes before a round number and 8 minutes after a round number). For example, to test for bunching at 3 hours and 30 minutes, we use a window from 3 hours and 22 minutes to 3 hours and 38 minutes. We choose this window to avoid bunching that may occur at a 10-minute mark in the counterfactual distribution either above or below the reference point of interest. We look for evidence of bunching itself in a 4-minute window right before each round number. As recommended by Chetty et al. (2011), this window was chosen based on visual inspection of the bunching. We employ a conservative test and use the same window for every potential reference point. Finally, before calculating the excess mass measure, we shift the entire counterfactual distribution upward so that the area underneath the counterfactual curve is equivalent to the area under the actual density function, thus avoiding the bias that would otherwise occur since the bunching is likely drawing from individuals just outside the bunching region. Thus, without this correction, we would essentially be double counting runners that are bunching at the reference point and causing the counterfactual distribution to be lower than it would otherwise be in a truly counterfactual world.

The main results of the bunching estimation applied to our full sample are depicted in Figure 3 and summarized in Table 2. Figure 3 graphically shows the 16-minute window around reference points at 3:00, 3:10, 3:20, 3:30, 4:00, 4:30, 5:00, and 6:00. The actual finishing times are plotted in 15-second intervals along with the counterfactual distribution that we estimate using the procedure above. The figures show clear evidence of bunching at the majority of the round number reference points. The bunching is particularly evident at the 3- and 4-hour marks. In Sections A.5–A.7 of the online appendix, we present the same results for all 10- and 15-minute marks from 2:30 to 6:00 and show that our results are robust to variations in the

Figure 3. Distribution of the Number of Finishers Around Round Number Reference Point and the Fitted Counterfactual Distribution



Notes. The vertical axis shows the number of finishers in each 15-second bin. The jagged line reflects the actual density function, and the smooth curve is the counterfactual density fitted by using the Chetty et al. (2011) procedure. The "bunching region" starts four minutes before a round number and ends at the round number.

16-minute window around a reference point, the 4-minute bunching region for excess mass, and other polynomial specifications for fitting the local density function.[13]

Table 2 provides summary measures from the procedure that is shown graphically in Figure 3. Specifically, we show the number of actual finishers in the 4-minute window around each of the round numbers as well as the number of finishers based on the counterfactual density function (after shifting the counterfactual function up). This gives us estimates for the number and percentage of excess finishers along with a t-statistic obtained by bootstrapping with 1,000 iterations. The largest number of runners (48,230) is displaced into the bunching region at four hours, whereas the largest percentage increase in finishers (24.2%) occurs at three hours. We find statistically significant evidence of bunching for all of the round numbers in Table 2 and, more generally, all 10-minute marks from 2:30 to 6:00, with the exception of 4:20, 4:50, 5:20, 5:40, and 5:50. This pattern of bunching also argues against left-digit bias as a mechanism. Left-digit bias is the

Table 2. Summary of Chetty et al. (2011) Test for Excess Mass

	Full sample (n = 9,789,093)				Nonmissing age and gender, 2002–2012 (n = 3,925,864)	Correcting for Boston marathon qualifier	
Reference point	Actual finishers	Counterfactual finishers	% excess finishers	t-statistic	% excess finishers	% excess finishers	t-statistic
3:00	90,762	73,077	24.2	52.95	23.7	23.7	27.82
3:10	115,770	109,077	6.1	15.76	7.7	4.2	4.95
3:20	165,968	164,131	1.1	3.59	1.6	0.9	1.46
3:30	259,756	234,405	10.8	42.09	11.1	11.1	21.90
4:00	419,945	371,715	13.0	64.34	13.3	13.4	35.50
4:30	316,967	303,231	4.5	20.09	4.7	4.7	12.47
5:00	218,170	206,858	5.5	19.14	5.6	5.6	12.96
5:30	115,737	113,314	2.1	5.78	2.8	2.8	5.24
6:00	63,643	61,694	3.2	6.12	3.0	3.0	4.42

Notes. t-Statistics are obtained using 1,000 bootstrap samples. The correction for the Boston Marathon uses the nonmissing age and gender/2002–2012 subsample (n = 3,925,864) and omits runners for which that reference point is a Boston Marathon qualifying time. For example, the 4:00 reference point omits males between the ages of 60 and 64 and females between the ages of 45 and 49.

tendency to focus more attention on the left-most digits of numbers than digits further to the right (Anderson and Simester 2003, Lacetera et al. 2012). Left-digit bias, however, provides an incomplete account of our bunching patterns. For example, left-digit bias predicts a similar amount of bunching at every left-digit change (3:10, 3:20, 3:30, 3:40, etc.). However, there is significantly more bunching at the rounder 3:30 and 4:30 marks than the less round 3:20, 3:40, 4:20, and 4:40 marks. Left-digit bias also cannot explain the small but statistically significant amount of bunching that occurs at 15-minute marks since there is no change in the left digit at those marks. Therefore, we argue that reference points established at round numbers offer a more natural psychological explanation of our pattern of data.

It is important to note that our instantiation of the Chetty et al. (2011) procedure is quite conservative. We use the same large bunching region for all tests to avoid issues with overfitting and to provide a measure of the number of excess finishers. Nevertheless, the panels in Figure 3 indicate that our bunching region provides significantly lower point estimates of the excess mass percentage, because it averages regions with considerable excess mass (the bins closest to the round number) with regions with less excess mass (the bins at the edge of the bunching region). For example, we find 24.2% excess mass ($t = 52.95$) at three hours using a bunching region of [2:56, 3:00]. The excess mass percentage is 28.3% ($t = 58.42$) if we restrict the bunching region to [2:57, 3:00] and 32.5% ($t = 45.64$) if the bunching region is [2:59, 3:00].

Finally, we examine the robustness of our bunching results by repeating the Chetty analysis for subsets of the data. These results, which use our full sample of data, are summarized in Table 3. We find that bunching of finishing times around the 3-, 4-, and 5-hour thresholds holds for recent marathons as well as marathons that took place decades ago; large as well as small marathons; relatively fast as well as relatively slow marathons; marathons in the United States, as well as marathons across other parts of the world; and, finally, for runners across a wide range of ages. Although the t-statistics naturally vary to reflect the different sample sizes for each data restriction, the effect sizes are remarkably uniform across different subsets of our marathon sample.

We also test for heterogeneity in bunching by running experience. Our measure of experience is the number of times a runner has previously run each given marathon. For example, a runner in the Chicago Marathon may have participated in three previous Chicago Marathons.[14] Previous research has found that experience may eliminate market anomalies such as the endowment effect (e.g., List 2003). However, experience in this domain does not provide a clear prediction. More experienced runners may be less likely to have reference-dependent preferences or nonround-number reference points; however, if they do, then they may be better at hitting targets/goals more easily than runners with less experience. Table 3 reports the amount of excess mass for marathon runners with varying levels of experience. The results do not suggest that there are any differences across the experience spectrum (the differences are small and not monotonic).

4.2. Boston Marathon Qualifying Times

We have suggested that our runners are reference dependent and that the bunching of finishing time is driven by the motivation to finish just ahead of a reference point. However, an alternative explanation for the bunching is that there is a change in the utility function at these round numbers due to an extrinsic benefit, as in Asch (1990). One obvious candidate for an extrinsic benefit at a round number is qualifying for the Boston Marathon. The Boston Marathon is the oldest annual marathon in the world and one of the few major marathons that has a qualifying time (although runners may also participate by working through charitable organizations). To qualify to participate in the Boston Marathon, a runner must complete a marathon faster than a qualifying time that is determined by that runner's age and gender. For example, from 2003 to 2012, the large majority of our sample, 18- to 34-year-old males, had to run a marathon in under 3 hours and 10 minutes to qualify for the Boston Marathon. The qualifying time for females of the same age was 3 hours and 40 minutes. Since the cutoffs for qualifying for the Boston Marathon are at round numbers, it is conceivable that this extrinsic reward is driving the observed bunching.[15]

It is fairly easy to show that the extrinsic benefit of qualifying to run the Boston Marathon cannot explain the full extent of our findings. For example, the 3-hour mark has not been a qualifying time for the Boston Marathon since 1989. Thus, bunching at 3 hours must be due to something else. Similarly, from 2003 to 2012, 4 hours only qualified 60- to 64-year-old males and 45- to 49-year-old females. Therefore, the bunching observed at the 4-hour mark must be driven by these two very small categories of runners. To systematically show how sensitive our results are to Boston Marathon qualifying times, we limit the sample to marathons conducted between 2002 and 2012 and include those whose age and gender indicate that the round number is not associated with a Boston Marathon qualifying time.[16] The last two columns in Table 1 indicate that these sample exclusions do very little to our estimates of excess finishers. The largest changes are at the 3:10 mark, where excess bunching drops from 7.7% to a still statistically significant 4.2%, and the 3:20 mark, where excess bunching drops from 1.6% to a statistically insignificant 0.9%.[17]

Table 3. Robustness Results of Excess Mass Measure for Subsets of Years, Number of Finishers, Mean Marathon Finishing Time, Geographical Region, and Age

				3:00 mark		4:00 mark		5:00 mark	
Data restriction	Number of marathons	Finishing time	Number of finishers	% excess finishers	t-statistic	% excess finishers	t-statistic	% excess finishers	t-statistic
Year									
≤1990	64	232.59	374,513	16.8	10.81	13.6	12.11	14.3	6.73
1991–2000	267	258.85	1,007,019	21.0	14.40	11.7	17.83	5.9	6.11
2001–2010	4,319	269.86	5,765,069	25.3	34.39	12.6	44.77	5.1	14.25
2011–2013	2,234	267.34	2,607,383	27.1	25.19	14.3	33.59	5.6	9.78
Number of finishers									
≥10,000	208	272.06	4,508,237	26.2	32.85	12.6	38.03	4.4	10.36
5,000–9,999	227	256.81	1,579,138	27.1	23.49	12.5	23.08	5.0	6.60
1,000–4,999	1,114	262.49	2,441,691	22.4	20.64	14.4	32.92	6.9	11.89
200–999	2,069	264.25	968,238	14.6	8.76	12.6	18.01	6.9	7.63
<200	3,270	275.75	291,789	14.8	4.63	8.8	6.65	8.3	4.96
Mean marathon finishing time									
≤4:00	703	230.75	1,305,498	25.4	28.30	13.5	25.01	7.1	5.66
(4:00,4:30]	3,619	256.68	5,358,117	24.7	34.85	13.6	47.74	5.2	13.58
>4:30	2,566	298.41	3,125,478	19.1	13.49	10.8	22.96	5.6	11.93
Marathon location									
United States	5,313	275.40	6,387,995	19.8	27.27	12.4	41.87	5.7	17.35
Europe	603	248.87	2,792,130	29.6	36.99	14.2	36.78	4.9	8.01
Canada	570	253.37	304,951	23.3	7.29	11.2	10.36	4.4	2.41
Other	402	256.20	304,017	21.9	8.87	14.3	11.92	5.7	3.30
Age									
≤29	—	275.57	1,168,315	21.9	13.13	13.1	19.27	3.9	5.19
30–39	—	267.30	1,733,982	23.4	19.13	13.5	24.79	5.2	7.84
40–49	—	267.76	1,562,812	27.3	19.25	13.4	24.58	6.2	9.28
≥50	—	285.98	1,027,637	27.7	9.89	12.3	16.54	6.1	7.82
Experience									
1st time	—	282.55	1,136,885	21.5	11.05	10.7	15.09	4.3	5.70
2nd time	—	276.05	272,591	24.1	6.79	11.4	8.42	6.0	3.57
3rd time	—	274.47	113,003	30.3	5.54	12.7	5.68	8.2	3.10
4th time	—	277.82	158,949	23.3	5.68	11.0	5.71	5.0	2.23

Notes. Analysis uses the total marathon sample ($n = 9{,}789{,}093$) with the exception of the age and experience restrictions. The data restrictions divide marathon-years by year, number of finishers, mean finishing time, and marathon location. In addition, we divide all participants by participant age and participant marathon experience as defined in Section 4.1. The excess mass measure uses the Chetty et al. (2011) procedure with the t-statistics obtained by using 1,000 bootstrap samples.

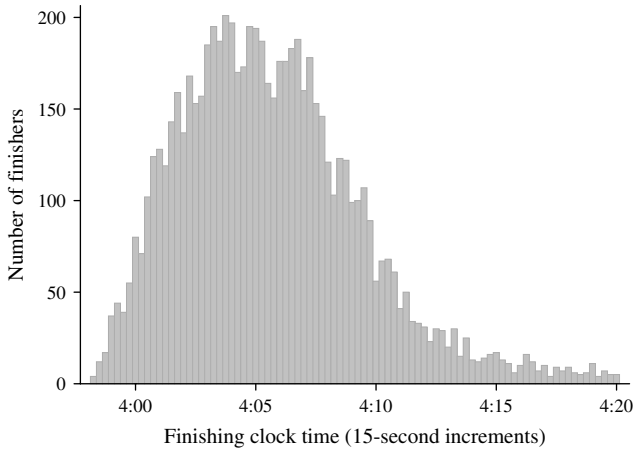
4.3. Pacesetters and Peer Effects

Most major marathons have pace teams to assist runners in achieving a desired time. For example, the 2013 Chicago Marathon provided pace teams for 3:00, 3:05, 3:10, 3:15, 3:20, 3:25, 3:30, 3:35, 3:40, 3:45, 3:50, 3:55, 4:00, 4:10, 4:25, 4:30, 4:40, 4:55, 5:00, 5:10, 5:25, and 5:45. The institution of pace teams then could provide an alternative explanation for the bunching we observe at round numbers. Although pacesetters are a reasonable alternative hypothesis, several pieces of evidence suggest that pacing cannot be the major driver of our effects. For example, the results we present in the next subsection on late race effort provision are difficult to explain with pacesetters. Additional evidence suggests that pacesetters cannot fully explain our results.

In large marathons such as the Chicago Marathon, runners cross the start line at very different clock times (the average difference between finishing clock time and finishing chip time in 2011 was 11.97 minutes). If pace teams are the primary explanation for our pattern of bunching, then a large group of runners should cross the finish line at the same clock time (since pacesetters can only work if runners are physically in the same area as the pacesetters). In Figure 4, we plot the finishing clock time for all Chicago Marathon runners with a chip time between 3:58 and 4:00. The figure shows that the clock times for these runners who bunch just short of 4 hours are very spread out. The fact that the runners who are bunching just before the 4-hour mark in chip time are finishing the race at very different clock times suggests that pacesetting is not a good explanation for the effects that we find.

A more direct way to rule out pacesetters as the primary driver of our results is to focus on marathons that

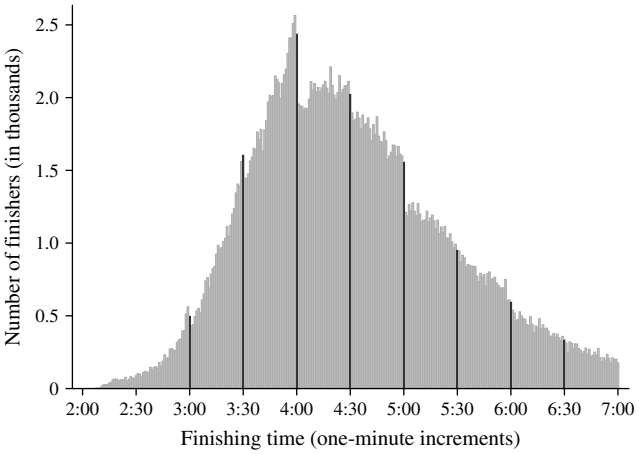
Figure 4. Distribution of Clock Times for Runners with a Chip Time Between 3:58 and 4:00, Chicago Marathon, 1998–2011



almost surely do not have institutionalized pace teams. We do so by examining small marathons. Figure 5 plots the distribution of finishing times for marathoners (n = 291,789) who participated in one of the 3,270 marathons with fewer than 200 finishers.[18] There continues to be strong graphical evidence of bunching at round numbers for these marathons with very few runners. Table 3 shows that the amount of excess mass for these small marathons is large and significant. Finally, formalized pace teams are a relatively new innovation, becoming widespread in the early 2000s, with the first instance in 1995.[19] Table 3 shows that bunching at the hour marks is substantial for marathons held before 1990 and between 1991 and 2000.

A related alternative explanation is that some of our bunching is driven by peer effects. In a classic study, Triplett (1898) found that cycling performance was facilitated by the presence of others. (Other economic analyses of peer effects are found in Falk and Ichino 2006, and Mas and Moretti 2009.) It is important to note that peer effects do not imply that there is no reference dependence but merely suggest that some of the bunching around round numbers might be driven by one marathoner running near another runner who has a round number time as a reference point. Our subset of small marathons also suggests that peer effects cannot be driving the bunching results, since the average difference in finishing times between one runner and the next runner is 2.61 minutes for all runners and 1.43 minutes for runners finishing between 3 hours and 50 minutes and 4 hours and 10 minutes.

Figure 5. Finishing Times for Marathons with Fewer than 200 Finishers



Note. The dark bars highlight the density in the 1-minute bin just before each 30-minute threshold.

4.4. How Does the Bunching Occur?

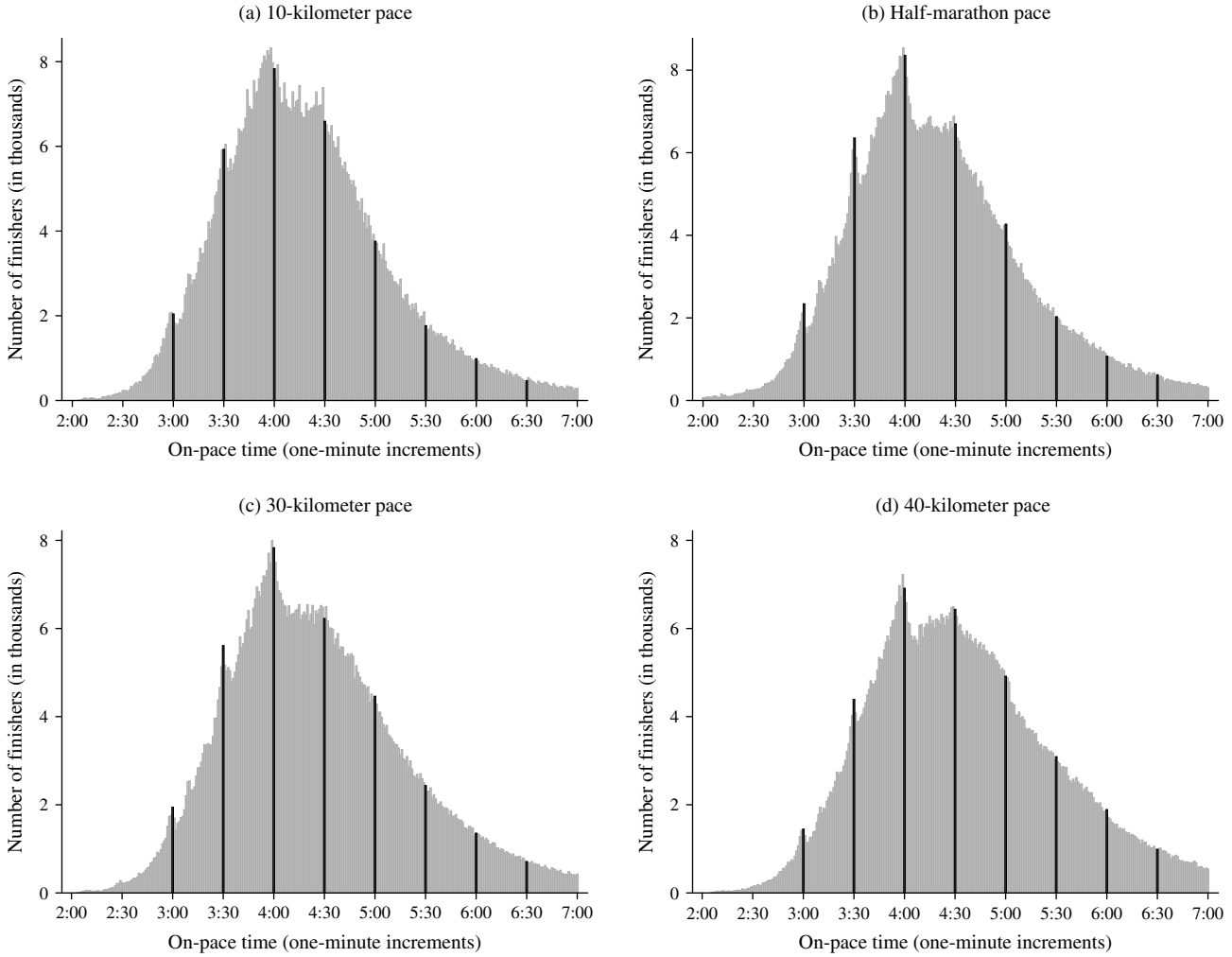
Individuals can respond to a kink in the tax code in several different ways. They can choose a job that pays an income close to the kink, plan their hours starting in January that will cause them to end at the kink, or adjust their hours in December in order to finish with income right at the kink.

Similarly, a marathon runner who has reference-dependent preferences can employ a number of different strategies that each could create bunching at the reference point. We use the richness of the marathon data to examine effort provision throughout various stages of the race. In particular, we explore two potential mechanisms for the bunching of finishing times. First, runners may adopt a reference point at the start of the race and pace themselves so as to finish just faster than that reference point. Second, runners may adjust their effort toward the end of the race so as to finish faster than a reference point. All of the analyses below are conducted on the full-split sample.

To look for evidence of reference-dependent pacing, we examine whether there is bunching in split times that correspond to a finishing time of a particular round number. For example, a 3-hour marathon is equivalent to a 10-kilometer split of 42 minutes and 40 seconds, a distinctly nonround number. Bunching of 10-kilometer-split times at 42.66 minutes would be evidence that runners are targeting a particular round number from the very beginning of the race.

Figure 6 shows the distribution of finishing times linearly extrapolated from split times at 10, 30, and 40 kilometers, as well as the half marathon. The bunching at split times equivalent to round number finishing times is not as stark as with actual finishing times at round numbers, but there is still clear evidence of bunching at each of these split times, with bunching becoming more pronounced as runners advance further in the marathon. For example, the excess mass percentages and t-statistics from the Chetty et al. (2011) analysis at the 3-hour marks are 6.7% and 4.89 (10 kilometers), 15.2% and 10.52 (half marathon), 12.2% and

Figure 6. Histogram of Extrapolated Finishing Times, Based on Intermediate Splits



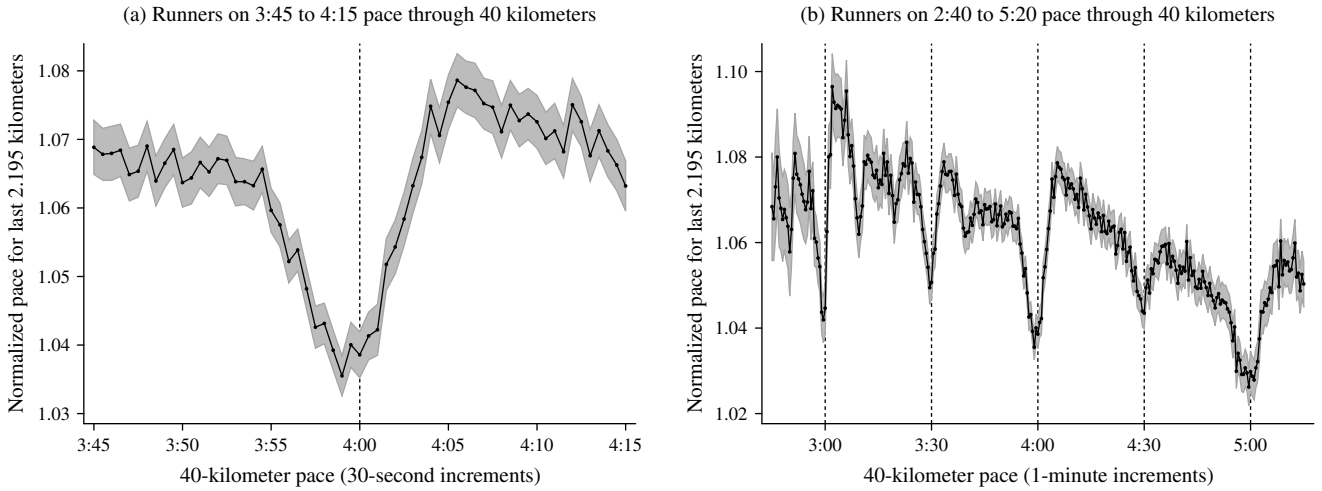
Notes. Splits times are extrapolated linearly to project finishing time. For example, the 10-kilometer split is multiplied by 4.2195.

8.15 (30 kilometers), 14.2% and 7.82 (40 kilometers), and 31.2% and 15.59 (finish). This analysis indicates that at least part of the bunching of finishing times is due to planning and pacing to better a round number reference point.

We next look for evidence that runners adjust their effort provision at the end of a race based on their proximity to a reference point. We start by calculating each runner's pace for the last 2.195 kilometers of the race relative to that runner's pace for the first 40 kilometers. We term this measure a runner's normalized pace, with a ratio of 1 indicating constant pace. A runner's 40-kilometer pace is calculated by multiplying the 40-kilometer split by 42.195/40.[20] Panel (a) of Figure 7 plots the normalized pace against the 40-kilometer pace, focusing on the runners who are on pace to finish in approximately 4 hours. The vertical axis in panel (a) indicates that runners ran the final 2.195 kilometers of their marathon on average 3%–8% slower than they ran the first 40 kilometers. However, normalized pace is clearly driven by a runner's pace through 40 kilometers. Runners who were on pace to finish between 3:45 and 3:55 or between 4:05 and 4:15 ran approximately 6%–8% slower in the last 2.195 kilometers. In contrast, runners who were on pace to finish close to the 4-hour mark (3:55–4:02) ran only 4%–6% slower in the last 2.195 kilometers. The sharp difference in relative pace as a function of proximity to the reference point is highlighted by the relatively narrow 95% confidence intervals. (We omit statistical tests because of the narrow confidence intervals.) Panel (b) of Figure 7 zooms out to show this normalized pace for runners across a wider range of 40-kilometer-pace times. The same qualitative pattern around 4 hours is observed at other round numbers in the distribution. Thus, there is clear evidence that runners finish the last 2.195 kilometers relatively faster when they are close to a round number than when they are farther away.

Note that this pattern of effort allocation is not because some runners choose different strategies for

Figure 7. Normalized Pace for Last 2.195 Kilometers as a Function of 40-Kilometer Pace



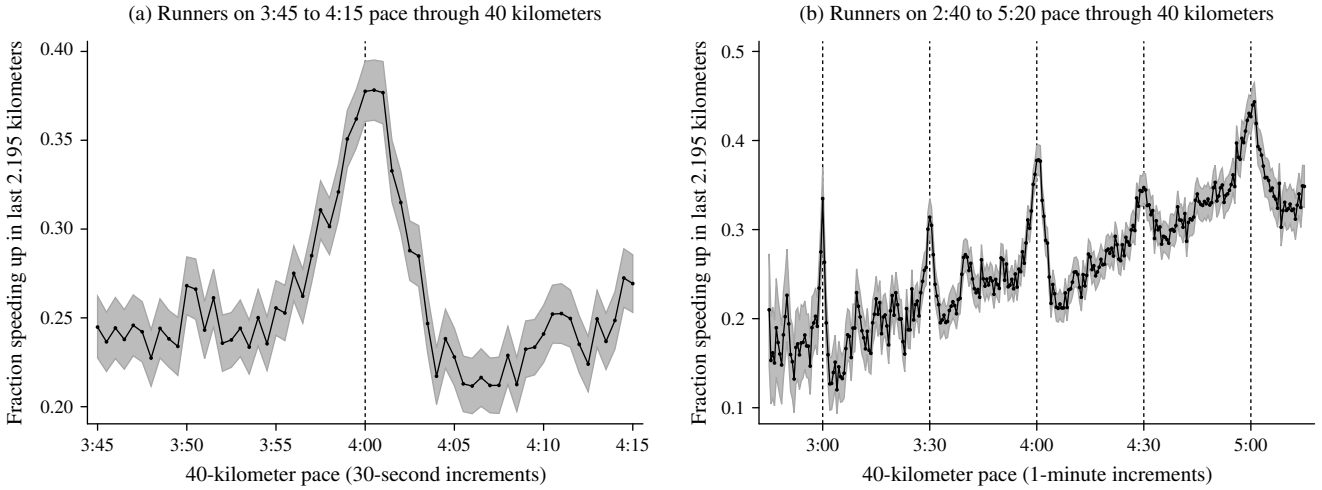
Notes. Normalized pace is calculated as the ratio of the pace for the last 2.195 kilometers (in minutes per kilometer) over the pace for the first 40 kilometers (also in minutes per kilometer). The plot shows normalized pace as a function of pace through 40 kilometers, linearly extrapolated to finishing time (i.e., the 40-kilometer split multiplied by 42.195/40); 95% confidence intervals are depicted by the shaded regions.

allocating energy. For example, the pattern could be explained by a mixture of runners who run more quickly from 30 to 40 kilometers, expending more effort so they can coast in, and runners who are more conservative, saving energy for a last push. A mixture of this kind would produce what looks like reference-dependent effort provision but would also result in a negative correlation between normalized pace from 30 to 40 kilometers and normalized pace from 40 kilometers on. On the contrary, the Spearman correlation between normalized pace from 30 to 40 kilometers and normalized pace over the last 2.195 kilometers is 0.54 ($p < 0.001$). The correlation remains positive ($\rho = 0.49$, $p < 0.001$) even when we drop the 25% of runners who slow down the most from 30 to 40 kilometers. Indeed, the pattern documented in Figure 7 holds if we normalize the pace for the last 2.195 kilometers relative to the pace from 30 to 40 kilometers.

Figure 7 shows that effort provision in the last 2.195 kilometers of a race depends heavily on a runner's proximity to a round number reference point. This speed adjustment can occur in different ways. For example, runners who are close to running 4 hours may be more likely to increase their speed relative to runners who are not near a round number. Alternatively, runners who are close to 4 hours may just be less likely to decrease their speed. We look at both speeding up and slowing down in Figures 8 and 9.

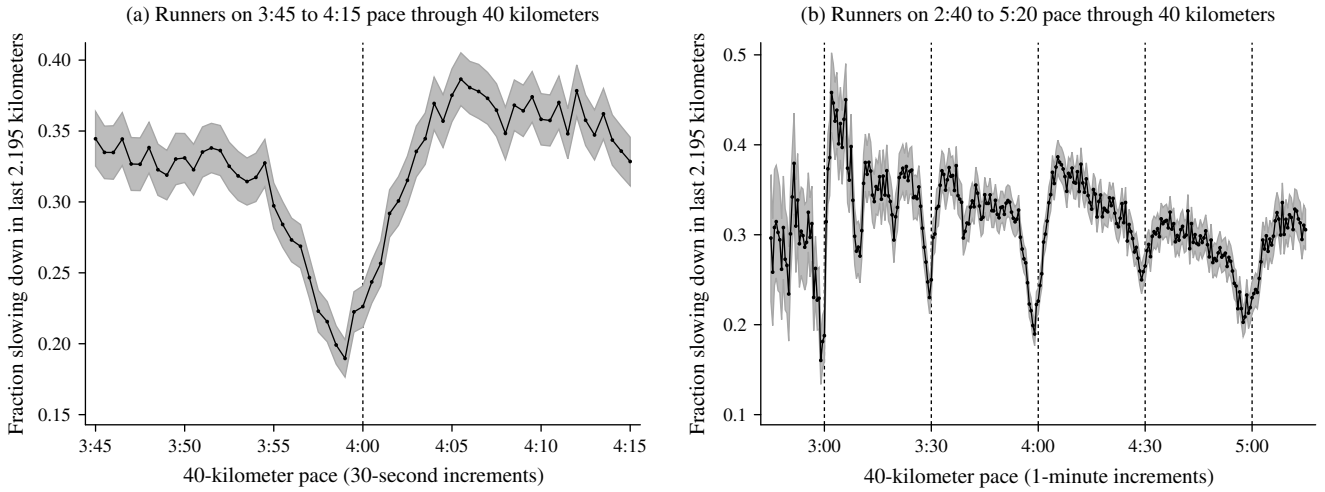
Figure 8 plots the probability that a runner runs the last 2.195 kilometers at a faster pace than the first

Figure 8. Percentage of Marathoners Who Speed Up Over the Last 2.195 Kilometers



Note. Speeding up is determined by comparing the pace for the last 2.195 kilometers to the pace for the first 40 kilometers; 95% confidence intervals are depicted by the shaded regions.

Figure 9. Percentage of Marathoners Who Slow Down over the Last 2.195 Kilometers by 10% or More



Note. Slowing down is determined by comparing the pace for the last 2.195 kilometers to the pace for the first 40 kilometers; 95% confidence intervals are depicted by the shaded regions.

40 kilometers. Panel (a) indicates that approximately 30% of runners increase their speed in the last 2.195 kilometers. This fraction, however, increases to almost 40% if a runner was right on target to finish at a round number. Once again, panel (b) shows the likelihood of speeding up across a wider range of 40-kilometer-pace times.

Most runners, however, are unable to maintain their pace near the end of the race. In fact, runners ran on average 5.6% slower over the last 2.195 kilometers. In Figure 9, we show the probability that a runner ran at least 10% slower over this last interval. Panel (a) depicts this probability near the 4-hour mark, whereas panel (b) looks at a wider range of 40-kilometer-pace times. We find that individuals who were just on pace to reach a round number were significantly less likely to slow down in the last stretch of the marathon than runners who were not in range to finish ahead of the reference point.

Collectively, Figures 6–9 indicate that finishing just short of a round number is driven by effort provision both in terms of planning and pacing, as well as the dynamic effort provision that occurs at the end of the marathon.

5. Discussion

We found significant bunching of marathon finishing times at round numbers. We hypothesized that this bunching was driven by reference dependence, as captured by models such as prospect theory, and showed that the stark bunching around the 30-minute marks is not caused by external benefits, such as qualifying for the Boston Marathon, or institutional features, such as pace groups. We proposed and found evidence for two mechanisms, planning and pacing and reference-dependent effort provision near the finish line. Although we do not report on these analyses here, we observe similar patterns for shorter racing distances, such as 10 miles and half marathons. However, these patterns are less pronounced, perhaps because these shorter races are run more often and thus reference points such as last or best previous performance are likely to substitute for round numbers.[21]

Our paper makes a number of broader contributions to the literature on reference dependence. First, our paper highlights the methodological benefits of nonparametric bunching estimation procedures for investigating patterns and implications of reference dependence. Unlike the studies of Camerer et al. (1997) and Crawford and Meng (2011), our procedure is nonparametric in that we do not estimate a reference point and instead allow the reference point to "pop out." This strategy is most clearly exploited in our analyses of effort provision toward the end of a marathon. This research clearly complements a growing body of research that follows similar strategies for identifying and estimating models of reference dependence. For example, Rees-Jones (2014) tested for evidence of loss aversion in tax sheltering and found significant bunching of tax returns around a zero balance. Baker et al. (2012) showed that stock offer price acceptances bunch at recent peak prices, consistent with recent peaks serving as reference points. Finally, DellaVigna et al. (2015) found bunching at unemployment insurance benefit cut points and argued that recent income serves as a reference point.

Second, our paper broadens the class of potential nonstatus-quo reference points. Much current work on reference dependence has taken either expectations or the status quo to be the reference point. The round number reference points in this investigation may be

what Rosch (1975) termed cognitive reference points. Round numbers (as well as focal colors) are cognitive reference points in the sense that stimuli are naturally compared to these reference points. Pope and Simonsohn (2011), for example, found that Major League Baseball players are more likely to finish the season with a 0.300 than a 0.299 batting average, and that high school students who take the SAT and just miss a round number score are more likely to retake the exam than those who just beat it. Because round numbers are so cognitively accessible, one clear organizational implication is that sales figures (or performance on many naturally occurring productivity tasks) may bunch around round numbers, even when these round numbers are not tied to extrinsic benefits (such as bonuses).

These round numbers may also be goals that individuals set to motivate themselves (Heath et al. 1999). Sackett et al. (2015) found that most marathoners set goals, but that these goals were for the most part optimistic, with 26% of runners achieving their self-reported goals. Thus, goals are clearly related to expectations, but unlike in the theoretical framework put forth by Kőszegi and Rabin (2006, 2007, 2009), goals are not always rational expectations. Thus, this paper also provides initial empirical linkages between a large psychological literature on the importance of goals (see Austin and Vancouver 1996 for a review of the psychology literature on goals, and Heath et al. 1999 for a psychological proposal that goals act as reference points) and emerging theoretical work in economics on goals and self-control (Hsiaw 2013, Koch and Nafziger 2011).

That said, in this setting, as in most natural field settings, other standards, besides round numbers, might also serve as reference points. We propose that there may be nothing special psychologically about round numbers relative to other reference points that a runner, or more generally an economic agent, might adopt. Put differently, we would expect empirical patterns similar to the ones we have documented in this paper to hold for nonround-number reference points. Round numbers do, however, play a unique and essential role in our empirical strategy. It is intuitive and indeed true that round numbers often serve as reference points, and, of course, we know when a number is in fact round.

Acknowledgments

The authors thank John List, the associate editor, and three reviewers for extremely useful comments. The authors are grateful for the comments of Jeffrey Allen, Han Bleichrodt, Colin Camerer, Stefano DellaVigna, Enrico Diecidue, David Erkens, Craig Fox, Alastair Lawrence, Maria Loumioti, Cade Massey, Pete McGraw, Canice Prendergast, Alex Rees-Jones, Richard Sloan, Nicholas Sly, Matthew Spiegel, and James Zuberi. The authors are also grateful for the comments of seminar participants at Tilburg University; University of California, Berkeley; University of Chicago; University of Oregon; University of Southern California; the Judgment and Decision Making (JDM) Winter Symposium; the Behavioral Economics Annual Meeting (BEAM); and the Natural Experiments Workshop in Munich, Germany. Ilknur Aliyev, Sandy Garcia, Dan Walco, Bing Wang, Jean (Jieyin) Zeng, and Rongchen Zhu provided invaluable research assistance. The authors also thank Dave McGillivray and Marc Davis of the Boston Athletic Association for providing historical data on Boston Marathon qualifying times.

Endnotes

[1] Section A.1 of the online appendix contains a list of field demonstrations of reference dependence.

[2] Temporal demonstrations of effort provision are relatively rare. See, however, Larkin (2014) and Misra and Nair (2011) for recent examples.

[3] Similarly, accounting research has documented bunching in firm financial performance: earnings, change in earnings, and earnings relative to analysts' consensus forecasts (e.g., Burgstahler and Dichev 1997, Hayn 1995). Do managers set these profit targets as goals (an intrinsic reference point) or are they concerned that investors, analysts, and the media are focused on these profit targets (an audience effect)?

[4] A similar argument could be made about football coaches who do not follow the optimal fourth-down strategy outlined by Romer (2006). Are these football coaches acting suboptimally, or are they merely reacting appropriately to fans, writers, and owners who are not sufficiently sophisticated to know that going for a first down is a better strategy than punting? Either way, someone, the coach or the audience, is making a mistake. See also Lefgren et al. (2015) for a similar point regarding outcome bias.

[5] The standard prospect theory value function is invariant to multiplicative scale. As a result, the notion of "more loss averse" could involve a "stretching" along the loss dimension, a "contraction" along the gain dimension, or both. Each interpretation yields bunching, although with different implications for whether the bunching comes from below (losses are stretched) or above (gains are contracted). Recent psychological (McGraw et al. 2010), measurement (Markle et al. 2015), and neuroimaging research (e.g., Tom et al. 2007) provides indirect but converging evidence for stretching of the scale in the loss domain, which we use as justification of this interpretation.

[6] Many of the marathons in our sample do not distinguish between chip time and clock time. In addition, the technology was not adopted by large marathons until 1996, when the Boston Marathon was the first U.S. marathon to use RFID chips to record marathon times (O'Connor 2007). When we only have a single finishing time as a measure of performance, we treat that time as if it were a chip time. Analyses reported in Endnote 13 indicate that this is a conservative assumption.

[7] Our data set comes from results posted on the websites of individual marathons and from http://www.marathonguide.com/, which has a relatively complete set of results for U.S. and Canadian (as well as some international) marathons from 2000 to the present. A full list of marathons in our sample is available at http://faculty.chicagobooth.edu/george.wu/research/marathon/list.htm.

[8] The considerably faster times in the full sample reflect a significantly older sample of marathon finishing times. In our sample, marathon finishing times have gotten slower by an average of 54 seconds each year.

[9] We created a panel data set of marathon finishing time by using names and ages as identifiers. We then tested whether a runner's previous marathon time served as a reference point for the subsequent

version of the same marathon. We found very limited evidence for bunching at this potential reference point.

[10] Figure A.1 in the online appendix contains a plot of the residuals between the actual density function and a 15th-order polynomial fitted to the density function in Figure 2.

[11] Note that all of these thresholds are to the left of 4:17:20, the median of the distribution, and thus this measure of excess mass would be negative for normal, lognormal and many other single-peaked continuous distributions.

[12] In Section A.4 of the online appendix, we conduct an alternative test in which we examine whether there is a significant discontinuity in the density function at round numbers and whether the largest discontinuity occurs at the round number, or merely around the round number. This test, which uses a regression discontinuity procedure developed by McCrary (2008), produces similar results to the analysis presented here.

[13] To verify that runners are using chip times and not clock times to evaluate their performance, we repeated the same analysis for clock time instead of chip time, finding considerably stronger results for chip time. To do so, we restricted our sample to runners with both clock and chip times (n = 5,618,168). Using the Chetty et al. (2011) procedure on this sample, we estimated 5.6% excess mass for clock time and 13.3% excess mass for chip time at four hours. The effect is even more dramatic when we restrict our analysis to runners with a clock time at least one minute slower than their chip time (n = 4,070,840) (2.9% excess mass for clock time and 13.1% excess mass for chip time) or at least two minutes slower than their chip time (n = 3,253,757) (1.0% excess mass for clock time and 12.6% excess mass for chip time). We find similar results at other round numbers.

[14] The data used for this heterogeneity cut are 11 of the 12 largest U.S. marathons (excluding Honolulu). For each of these marathons, we have the first and last name of each runner. We restrict the sample to names that are not so common as to show up multiple times in the same marathon (this name restriction drops 8% of the sample). We also restrict the sample to marathons run after 2003 (the years before 2003 are used to help to calculate the number of marathons a runner has run as of 2003 and beyond). Based on this sample, we are able to construct a measure for each observation that indicates how many times the runner previously ran a given marathon.

[15] From 1997 to 2012, the Boston Marathon rounded times down, and thus a time of 3:10:59 qualified a 31-year-old male runner. This threshold suggests that we should find bunching at 3:11, rather than 3:10. In contrast, we observe considerably more excess mass for [3:06, 3:10] (6.1%) than for [3:07, 3:11] (3.9%).

[16] To estimate these new effects, we restrict the data to marathons for which we have both the age and gender for each runner. The third-to-last column in Table 2 replicates our earlier results using this restricted sample and serves as a baseline to compare the Boston Marathon qualifying results.

[17] Some marathons have limited capacity, with demand for entries outstripping supply. Although entry is often determined by lottery, some marathons provide a "guaranteed entry" based on time qualification. In most cases these standards are more challenging than the Boston Marathon qualifying standards. For example, the standards for the Berlin Marathon are 2:45 for men 45 and under, and 3:00 for women 45 and under. The New York Marathon standards are almost as difficult, with a qualifying time of 2:53 for males 18–34. Note that 8 of the 13 qualifying times under 4 hours for New York occur at nonround numbers such as 2:53. Qualifying times for the Chicago Marathon are 3:15 for men and 3:45 for women. Although the Houston Marathon has a 4-hour qualifying time for men and women, the relatively small size of this marathon (6,664 finishers in 2013), the difficulty of achieving most guaranteed entry times, and the robustness of our results across round number levels, locations, and years suggests that these standards cannot fully explain the observed patterns of bunching.

[18] Indeed, the website http://findmymarathon.com/ indicates that none of these marathons have pace teams.

[19] http://runcim.org/got-pacers-you-bet/ (accessed April 10, 2016).

[20] Because the split data are somewhat noisy (almost 5% of the results have a slower 40-kilometer split than finishing time), we excluded the bottom and top 5% of normalized pace data.

[21] Grant (2014) shows how ultramarathoners exert effort to finish a 100-mile race within 24 hours.

References

Anderson ET, Simester D (2003) Effects of $9 price endings on retail sales: Evidence from field experiments. *Quant. Marketing Econom.* 1(1):93–110.

Asch BJ (1990) Do incentives matter? The case of Navy recruiters. *Indust. Labor Relations Rev.* 43(3):89S–106S.

Ashenfelter O, Doran K, Schaller B (2010) A shred of credible evidence on the long-run elasticity of labour supply. *Economica* 77(308):637–650.

Austin JT, Vancouver JB (1996) Goal constructs in psychology: Structure, process, and content. *Psych. Bull.* 120(3):338–375.

Baker M, Pan X, Wurgler J (2012) The effect of reference point prices on mergers and acquisitions. *J. Financial Econom.* 106(1):49–71.

Barberis NC (2013) Thirty years of prospect theory in economics: A review and assessment. *J. Econom. Perspect.* 27(1):173–196.

Burgstahler D, Dichev I (1997) Earnings management to avoid earnings decreases and losses. *J. Accounting Econom.* 24(1):99–126.

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

Card D, Mas A, Moretti E, Saez E (2012) Inequality at work: The effect of peer salaries on job satisfaction. *Amer. Econom. Rev.* 102(6):2981–3003.

Chetty R, Friedman JN, Olsen T, Pistaferri L (2011) Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from Danish tax records. *Quart. J. Econom.* 126(2): 749–804.

Crawford VP, Meng J (2011) New York City cab drivers' labor supply revisited: Reference-dependent preferences with rational-expectations targets for hours and income. *Amer. Econom. Rev.* 101(5):1912–1932.

DellaVigna S (2009) Psychology and economics: Evidence from the field. *J. Econom. Literature* 47(2):315–372.

DellaVigna S, Lindner A, Reizer B, Schmieder JF (2015) Reference-dependent job search: Evidence from Hungary. Working paper, University of California, Berkeley.

Diecidue E, Van De Ven J (2008) Aspiration level, probability of success and failure, and expected utility. *Internat. Econom. Rev.* 49(2):683–700.

Falk A, Ichino A (2006) Clean evidence on peer effects. *J. Labor Econom.* 24(1):39–57.

Farber HS (2005) Is tomorrow another day? The labor supply of New York City cabdrivers. *J. Political Econom.* 113(1):46–82.

Farber HS (2008) Reference-dependent preferences and labor supply: The case of New York City taxi drivers. *Amer. Econom. Rev.* 98(3):1069–1082.

Farber HS (2015) Why you can't find a taxi in the rain and other labor supply lessons from cab drivers. *Quart. J. Econom.* 130(4): 1975–2026.

Fehr E, Goette L (2007) Do workers work more if wages are high? Evidence from a randomized field experiment. *Amer. Econom. Rev.* 97(1):298–317.

Fishburn PC (1977) Mean-risk analysis with risk associated with below-target returns. *Amer. Econom. Rev.* 67(2):116–126.

Grant DP (2014) The essential economics of thresholds: Theory and estimation. Working paper, Sam Houston State University, Huntsville, TX.

Hayn C (1995) The information content of losses. *J. Accounting Econom.* 20(2):125–153.

Heath C, Larrick RP, Wu G (1999) Goals as reference points. *Cognitive Psych.* 38(1):79–109.

Hsiaw A (2013) Goal-setting and self-control. *J. Econom. Theory* 148(2):601–626.

Kahneman D (1992) Reference points, anchors, norms, and mixed feelings. *Organ. Behav. Human Decision Processes* 51(2):296–312.

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263–291.

Kleven HJ, Waseem M (2013) Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan. *Quart. J. Econom.* 128(2):669–723.

Koch AK, Nafziger J (2011) Self-regulation through goal setting. *Scandinavian J. Econom.* 113(1):212–227.

Kőszegi B, Rabin M (2006) A model of reference-dependent preferences. *Quart. J. Econom.* 121(4):1133–1165.

Kőszegi B, Rabin M (2007) Reference-dependent risk attitudes. *Amer. Econom. Rev.* 97(4):1047–1073.

Kőszegi B, Rabin M (2009) Reference-dependent consumption plans. *Amer. Econom. Rev.* 99(3):909–936.

Lacetera N, Pope DG, Sydnor JR (2012) Heuristic thinking and limited attention in the car market. *Amer. Econom. Rev.* 102(5): 2206–2236.

Larkin I (2014) The cost of high-powered incentives: Employee gaming in enterprise software sales. *J. Labor Econom.* 32(2):199–227.

Lefgren L, Platt B, Price J (2015) Sticking with what (barely) worked: A test of outcome bias. *Management Sci.* 61(5):1121–1136.

List JA (2003) Does market experience eliminate market anomalies? *Quart. J. Econom.* 118(1):41–71.

March J, Shapira Z (1987) Managerial perspectives on risk and risk taking. *Management Sci.* 33(11):1404–1418.

Markle AB, Wu G, White RJ, Sackett AM (2015) Goals as reference points in Marathon running: A novel test of reference-dependence. Working paper, Fordham University, New York

Mas A (2006) Pay, reference points, and police performance. *Quart. J. Econom.* 121(3):783–821.

Mas A, Moretti E (2009) Peers at work. *Amer. Econom. Rev.* 99(1): 112–145.

McCrary J (2008) Manipulation of the running variable in the regression discontinuity design: A density test. *J. Econometrics* 142(2):698–714.

McGraw AP, Larsen JT, Kahneman D, Schkade D (2010) Comparing gains and losses. *Psych. Sci.* 21(10):1438–1445.

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

Murphy KJ (2000) Performance standards in incentive contracts. *J. Accounting Econom.* 30(3):245–278.

O'Connor F (2007) RFID helps the Boston Marathon run. *PC World* (April 9), http://www.washingtonpost.com/wp-dyn/content/article/2007/04/09/AR2007040901011.html.

Oyer P (1998) Fiscal year ends and nonlinear incentive contracts: The effect on business seasonality. *Quart. J. Econom.* 113(1): 149–185.

Pope D, Simonsohn U (2011) Round numbers as goals. *Psych. Sci.* 22(1):71–79.

Prendergast C (1999) The provision of incentives in firms. *J. Econom. Literature* 37(1):7–63.

Rees-Jones A (2014) Loss aversion motivates tax sheltering: Evidence from U.S. tax returns. Working paper, University of Pennsylvania, Philadelphia.

Romer D (2006) Do firms maximize? Evidence from professional football. *J. Political Econom.* 114(2):340–365.

Rosch E (1975) Cognitive reference points. *Cognitive Psych.* 7(4): 532–547.

Running USA (2014) Running USA's 2014 annual marathon report. (March 23), http://www.runningusa.org/index.cfm?fuseaction=news.details&ArticleId=332.

Sackett AM, Wu G, White RJ, Markle AB (2015) Harnessing optimism: How eliciting goals improves performance. Working paper, University of St. Thomas, St. Paul, MN.

Saez E (2010) Do taxpayers bunch at kink points? *Amer. Econom J.: Econom. Policy* 2(3):180–212.

Tom SM, Fox CR, Trepel C, Poldrack RA (2007) The neural basis of loss aversion in decision-making under risk. *Science* 315(5811):515–518.

Triplett N (1898) The dynamogenic factors in pacemaking and competition. *Amer. J. Psych.* 9(4):507–533.

Tversky A, Kahneman D (1991) Loss aversion in riskless choice: A reference dependent model. *Quart. J. Econom.* 106(4):1039–1061.

Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertainty* 5(4): 297–323.

Wakker PP, Tversky A (1993) An axiomatization of cumulative prospect theory. *J. Risk Uncertainty* 7(2):147–176.