NBER WORKING PAPER SERIES

REFERENCE-DEPENDENT PREFERENCES:
EVIDENCE FROM MARATHON RUNNERS

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

Working Paper 20343
http://www.nber.org/papers/w20343

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2014

We 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, Richard Sloan, Nicholas Sly, Matthew Spiegel, James Zuberi, and seminar participants at Tilburg University, University of California, Berkeley, University of Chicago, University of Oregon, and University of Southern California and at the 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, and Rongchen Zhu provided invaluable research assistance. We also thank Dave McGillivray and Marc Davis of the Boston Athletic Association for providing us with historical data on Boston Marathon qualifying times. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Reference-Dependent Preferences: Evidence from Marathon Runners
Eric J. Allen, Patricia M. Dechow, Devin G. Pope, and George Wu
NBER Working Paper No. 20343
July 2014
JEL No. D03,J22

ABSTRACT

Models 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 dataset of marathon finishing times (n = 9,524,071). 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). We calibrate a simple model of prospect theory as well as other models of reference dependence and show that the basic qualitative shape of the empirical distribution of finishing times is consistent with parameters that have previously been estimated in the laboratory.

Eric J. Allen
USC Marshall School of Business
Los Angeles, CA 90089-0808
ericalle@marshall.usc.edu

Patricia M. Dechow
Haas Accounting Group
University of California at Berkeley
Berkeley, CA
patricia_dechow@haas.berkeley.edu

Devin G. Pope
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
devin.pope@chicagobooth.edu

George Wu
Booth School of Business
University of Chicago
Chicago, IL
wu@chicagobooth.edu

1 Introduction

Recent theories of economic behavior propose that preferences are reference dependent. In other words, 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, Mas, Moretti, and Saez, 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. For example, a primary feature of prospect theory, the most well-known and influential account of reference-dependent preferences, is that the first derivative of utility is discontinuous at the reference point (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992). This property, known as *loss aversion*, has implications for a number 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*, results in a discontinuous second derivative at the reference point and is captured by prospect theory's characteristic S-shaped value function that is concave for gains and convex for losses. While 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. More generally, we suggest that the distinguishing feature of reference-dependent models is a discontinuity at the reference point that is psychologically-based and not the result of some extrinsic benefit.

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

[1] In Finance, prospect theory has shed light on the equity premium puzzle (Benartzi and Thaler, 1995), the disposition effect (Odean, 1998; Shefrin and Statman, 1985), and stock option exercise decisions (Heath, Huddart, and Lang, 1999). Barberis and Thaler (2003) provide a survey of the behavioral finance literature that offers a thorough discussion of prospect theory and the impact that it has had in finance. Prospect theory is also tied closely with work on the endowment effect and status-quo bias (e.g., Kahneman, Knetsch and Thaler, 1990; Samuelson and Zeckhauser, 1988). Other important applications of prospect theory include consumer behavior (Hardie, Johnson, and Fader, 1993), housing decisions (Genesove and Mayer, 2001), international trade (Tovar, 2009), game shows (Post, Van den Assem, Baltussen, and Thaler, 2008), insurance (Barseghyan, Molinari, O'Donoghue, and Teitelbaum, 2013; Sydnor, 2010), sports (Berger and Pope, 2011; Pope and Schweitzer, 2011), framing and social comparisons in organizations (Card, Mas, Moretti, and Saez, 2012; Hossain and List, 2012), education (Fryer, Levitt, List, and Sadoff, 2012; Levitt, List, Neckerman, and Sadoff, 2012), tax deductions (Rees-Jones, 2013), gambling (Lien, 2013), and household behavior (Bertrand, Pan, and Kamenica, 2013). Recent reviews of applied work in behavioral economics that discuss this literature include DellaVigna (2009), Barberis (2013), and Pope and Sydnor (forthcoming).

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 due to reference-dependent preferences defined by daily income targets (see also Fehr and Goette, 2007, and Mas, 2006). This paper led to additional analyses that used different datasets and econometric methods to examine if taxi drivers indeed have reference-dependent preferences, with some arguing against (Farber, 2005, 2008) and some arguing in favor (Ashenfelter, Doran, and Schaller, 2010; Crawford and Meng, 2011). The primary empirical challenge in these papers has been modeling reference points that are unobservable, heterogeneous, and possibly non-stationary.

In this paper, we test for reference dependence in a dataset of over 9 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 provide survey evidence that the majority of runners think about their performance relative to round numbers (e.g., running a marathon in 4 hours). The prevalence across runners of round number reference points provides us with a much cleaner and sharper test of reference dependence than other settings where reference points are unknown or likely to differ across individuals. Coupled with our large sample, these universal reference points allow us to very easily and credibly identify evidence of reference dependence using non-parametric methods. We also exploit the richness of our data to directly examine how reference points impact effort provision at different points in the race.[2]

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, 51.4% more runners finished in the minute just under 3 hours than the minute just over 3 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., pace setters) and show that this effect is explained in part by pacing and planning and in part by effort provision over the final 2.195 kilometers of the marathon. Runners are more likely to speed up and less likely to slow down if they are on pace to finish just ahead of a round number reference point. We also show that runners who start the race with one reference point may nevertheless spontaneously adopt a slower round number reference point once they realize that they are unable to reach their original target.

[2] Temporal demonstrations of effort provision are relatively rare. See, Larkin (2014) for a recent example in a pay-for-performance context.

We perform a calibration exercise to better understand what utility functions can produce the type of bunching exhibited in marathon finishing times. A utility function with diminishing sensitivity and either a notch or a kink at the reference point can produce the pattern of bunching exhibited in marathon finishing times. In the second case, we find that our results are roughly consistent with standard prospect theory parameter estimates.

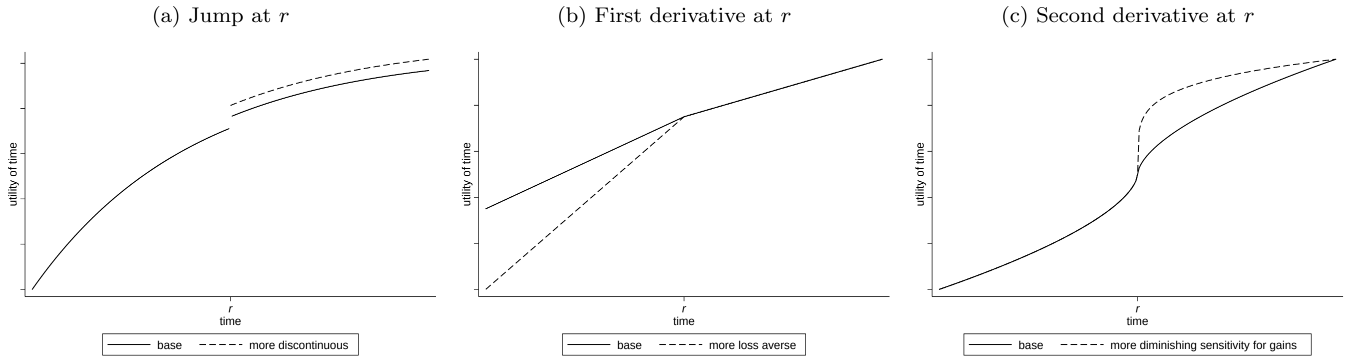
In addition to providing a particularly compelling and clean test for reference dependence, our paper makes a conceptual contribution. In our setting, we argue that the relevant reference points are goals. This research is thus set apart from most current work on reference dependence which takes either expectations or the status quo to be the reference point. Goals are clearly related to expectations, but unlike the theoretical framework put forth by Kőszegi and Rabin (2006, 2007, 2009), goals are not rational expectations. For example, only 26% of participants in Sackett, Wu, White, and Markle's (2014) study of marathon runners achieved their self-reported goals. This paper thus broadens the class of potential non-status quo reference points. It also provides 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, Larrick, and Wu's (1999) 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).[3]

Another paper that has many similarities to this one is Pope and Simonsohn (2011), who argued that round numbers can act as goals. They found that Major League Baseball players are more likely to finish the season with a 0.300 batting average as opposed to 0.299, 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. In contrast to that paper, we focus on how round numbers can serve as reference points in a setting that is largely void of extrinsic financial or other direct incentives. Performance on professional baseball statistics and college admissions exams are tied very directly to future financial compensation. As a result, there is the concern that Pope and Simonsohn's non-laboratory findings could be attributable to responses to financial incentives as opposed to internally-generated reference points (Moskowitz and Wertheim, 2011).

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 that taxpayers may 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

[3] Our paper is also related to ongoing work focused on goals and marathon performance. Sackett, Wu, White, and Markle (2014) conducted a field experiment on marathon runners. In one treatment, runners were asked to provide a goal prior to the marathon. In a control treatment, runners merely provided demographic information. The first treatment did not influence the likelihood that runners had goals, but did lead experienced runners to set significantly more ambitious goals and run significantly faster. In a related paper, Markle, Wu, White, and Sackett (2014) related self-reported satisfaction and performance. See Section 2 for more detail.

Figure 1: Three forms of reference-dependent preferences


(a) Jump at r

(b) First derivative at r

(c) Second derivative at r

Section 4. In Section 5, we calibrate a simple model of prospect theory as well as a reference-dependent model with a jump at the reference point and show that previous prospect theory parameter estimates can reproduce the pattern and magnitude of bunching found in our archival data. We conclude the paper in Section 6 with a brief discussion of the broader significance of our findings.

2 Conceptual Framework

In this section, 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 qualitiatively distinct perception of outcomes that are just above and 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: 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 might be referred to as 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 (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 (Kőbberling and Wakker, 2005; Tversky and Kahneman, 1992). The solid lines in Figure 1 shows discontinuities of all three forms. Note that Markle, Wu, White, and Sackett (2014) found evidence for all three of these forms when they related marathoners' performance and self-reported satisfaction. Satisfaction as a function of performance relative to a goal exhibited loss aversion, diminishing sensitivity, and 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. 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., $U(t) = b(t) - c(t)$. Furthermore, we assume that $c(t) > 0$ and $c'(t) > 0$, i.e., costs are positive and increasing. We also assume that the benefit function has at least one of the three forms of reference-dependent preferences described above. Throughout, we take agents to be optimizers, choosing t to maximize $U(t)$. We denote $t^*\big(c(t), b(t)\big)$ 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 statistics 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, $C\big(\delta, b(t)\big)$, exhibits δ-bunching around reference point r if, for all $c(t) \in C\big(\delta, b(t)\big)$, $0 \leq t^*(c(t), b(t)) - r \leq \delta$.

It is clear that a jump or notch at the reference point will lead to bunching just past 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.

Definition (More Discontinuous at r) A benefit function $b_1(t)$ is more discontinuous at reference point r than $b_2(t)$ if $\lim_{\epsilon \to 0} b_1(r+\epsilon) - \lim_{\epsilon \to 0} b_1(r-\epsilon) > \lim_{\epsilon \to 0} b_2(r+\epsilon) - \lim_{\epsilon \to 0} b_2(r-\epsilon)$.

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

The proof is straightforward and omitted. Here, a psychological jump in utility acts produces behavior 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 $b'(r+\epsilon) < b'(r-\epsilon)$ 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 $b_1(t)$ and $b_2(t)$ both exhibit loss aversion, and $b_1(t) = b_2(t)$ 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 δ, $C\big(\delta, b_2(t)\big) \subseteq C\big(\delta, b_1(t)\big)$.*

We interpret Proposition 2.2 as indicating that as loss aversion increases, more individuals will bunch up 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.[4]

[4] The standard prospect theory value function is invariant to multiplicative scale. Thus, in this context, the notion of more loss averse is ambiguous, because it may involve a "stretching" along the loss dimension, a "contraction" along the gain dimension, or both. Each interpretation yields bunching, although the different interpretations have different implications for whether the bunching comes from below (losses are stretched) or above (gains are contracted). Recent psychological (McGraw, Larsen, Kahneman, and Schkade, 2010), measurement (Markle, Wu, White, and Sackett, 2014), and neuro research (e.g., Tom, Fox, Trepel, and Poldrack, 2007) provides indirect but converging evidence for stretching of the scale in the loss domain, which we use as justification of this interpretation.

Proof We prove the result by contradiction. Assume that Proposition 2.2 is false. Then there exists some $\hat{c}(t)$ such that $\hat{c}(t) \in C\big(\delta, b_2(t)\big)$, but $\hat{c}(t) \notin C\big(\delta, b_1(t)\big)$. If $\hat{c}(t) \notin C\big(\delta, b_1(t)\big)$, then either $t^*(\hat{c}(t), b_1(t)) - r > \delta$ or $t^*\big(\hat{c}(t), b_1(t)\big) - r < 0$. However, the first inequality contradicts $b_1(t) = b_2(t)$ for $t > r$ and the second inequality contradicts $b_1'(-t) < b_2'(-t)$ for $-t < r$. ■

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 on $(r, r+\gamma)$ if $b_1''(t) < 0$ and $b_2''(t) < 0$ for $t > r$, and $b_1(t) = f\big(b_2(t)\big)$ for $t \in (r, r+\gamma)$ and $b_1(t) = b_2(t)$ otherwise, where $f(\cdot)$ is a strictly concave function.

See Figure 1, Panel C, 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 $(r, r+\gamma)$. In that interval, $b_1(t)$ is strictly more concave than $b_2(t)$. Note that the Proposition requires that $b_1(r+\gamma) = b_2(r+\gamma)$ so that the cumulative benefits on $[r, r+\gamma]$ 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 $(r, r+\gamma)$. Then $C\big(\gamma, b_2(t)\big) \subseteq C\big(\gamma, b_1(t)\big)$.*

Proof We prove the result by contradiction. Assume that Proposition 2.3 is false. Then there exists some $\hat{c}(t)$ such that $\hat{c}(t) \in C\big(\gamma, b_2(t)\big)$, but $\hat{c}(t) \notin C\big(\gamma, b_1(t)\big)$. This holds if $b_1'(r+\gamma) > \hat{c}'(r+\gamma) > b_2'(r+\gamma)$. Since $b_1(r+\gamma) = f\big(b_2(r+\gamma)\big)$, by the chain rule, $b_1'(r+\gamma) = f'\big(b_2(r+\gamma)\big)b_2'(r+\gamma)$. Thus $b_1'(r+\gamma) > b_2'(r+\gamma)$ holds if $f'(b_2(r+\gamma)\big) > 1$. However, if $f'\big(b_2(r+\gamma)\big) > 1$ and $f'(t)$ is a strictly concave function, then $f'\big(b_2(t)\big) > 1$ for $r \leq t \leq r+\gamma$ and therefore $b_1(r+\gamma) = f\big(b_2(r+\gamma)\big) > b_2(r+\gamma)$, which is a contradiction. ■

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.

It is important to note that our simple model does not involve any risk preferences. Of course, in a marathon setting, it is unlikely that a runner knows his or her actual cost function on a particular day. Although incorporating risk preferences into this framework will produce similar comparative statics, a model with uncertainty will clearly have implications for the specific shape of the finishing time density function. We revisit this complication in our discussion of calibration in Section 5.

The conceptual framework we have laid out in this section illustrates 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 will produce bunching or excess mass at reference points, even if the family of cost functions is smooth. We further proved that the amount of bunching that we predict is 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 and in Section 5 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 kilometers (26.2 miles) road race that is popular with both professional athletes and recreational runners. Approximately 1,100 marathons were held in the U.S. during 2013, with an estimated 541,000 finishers.[5] The vast majority of runners receive no financial compensation for their performance. For example, in 2013, the 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.

For our purposes, an important technological innovation in marathon running is a radio frequency (RFID) chip that is attached to a runner's shoelace or running bib. This chip precisely measures a runner's finishing time. For large marathons, many runners do not cross the starting line until many minutes after the official start (e.g., it took runners in the 2011 Chicago Marathon an average of 11.97 minutes to reach the start line). The computer chip registers when a runner crosses the starting line, the finishing line, and various intermediate points on the course (often 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, while 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

[5] http://www.runningusa.org/marathon-report-2014?returnTo=annual-reports

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 is therefore essential for our purposes.[6]

The data used in this paper were obtained from various public websites.[7] In total, we have complete finishing times for 9,524,071 marathon finishes. The full sample contains data from 1970-2013 (91.23% is 2000 or later) for 6,831 different marathon-years. Our full sample contains mostly U.S. marathons, but also includes the five largest Canadian marathons, and several large marathons from Europe, South America, Africa, Asia, and Australia. Our marathon sample includes multiple years of all of the 51 largest U.S. marathons (as measured by 2011 rank), as well as a relatively complete sample of all U.S. and Canadian marathons from 2000 to 2013. We will show results using this full sample, but for some of our analysis, we will focus on a smaller sample of 868,039 finishing times with complete 10 kilometer, half marathon, 30 kilometer, and 40 kilometer split times in order to make effect sizes comparable. 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 27 minutes and 00 seconds (4:27:00 for short) for the full sample and 4:42:07 seconds for the full split sample. The average half marathon split time is 2:11:55 in the full split sample, indicating that runners typically run faster in the first half of the marathon than in the second half.

There are many potential reference points that a runner may use for judging his or her marathon performance. For example, it is possible that a runner contrasts 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.[8] In this paper, we focus on a particular set of reference points that might affect runners: round numbers (e.g., 4-hour marathon time). There are two primary reasons that we focus our analyses on these reference points. First, these reference points are knowable to us as researchers. While the finishing time of a close friend might be an important reference point for some runners, we are unable to test for evidence of this due to data limitations. Second,

[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. When we only have a single finishing time as a measure of performance, we treat that time as if it was a chip time. Analyses reported in footnote 13 indicate that this is a conservative assumption.

[7] Our dataset comes from results posted on the websites of individual marathons, as well as from `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] We created a panel dataset 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.

round numbers are frequently mentioned as goals by marathon runners themselves. For example, Sackett, Wu, White, and Markle (2014) asked marathoners running 15 major U.S. marathons from 2007 to 2009 to provide their specific time goal. 86.3% of marathoners in that study indicated that they had a time goal. Of these individuals, 27.8%, 49.4%, and 70.5% had time goals that were divisible by 60, 30, 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.

The institutional details described in this section suggest that marathons are an ideal setting to look for evidence of internal reference points. Most notably, the reference points that many runners adopt are known to the researcher and are not directly tied to financial rewards. These facts allow us to test in a very direct way whether runners evaluate their finishing times relative to seemingly irrelevant numbers. However, while marathon finishing times are likely to be evaluated primarily by the runner, runners may also care about how their finishing time is perceived by others. If this is the case, evidence for reference dependence may not reflect internal reference points but instead be due to 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 as opposed to a 4:01 marathon (an internal reference point), or 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 internal versus audience effects. For example, does the taxi driver care about reaching a particular target, or are they worried about their 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 he is worried about what neighbors and family members will think if they discover that he lost money on this transaction? We will refer to the results found in the next section as evidence for internal reference points, but acknowledge that part of the effect may also be driven by others who similarly evaluate outcomes relative to round numbers. 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 (i.e., internal reference points) or with someone else (i.e., an audience effect).[9]

Another alternative mechanism for bunching of finishing times could be left-digit bias (Anderson and Simester, 2003; Lacetera, Pope, and Sydnor, 2012). Left-digit bias is the tendency of individuals to focus more attention to the left-most digits of numbers than digits further to the right. This bias has been used to

[9] A similar argument could be made for football coaches who do not follow the optimal fourth-down strategy outlined by Romer (2006). Are these football coaches making a mistake, 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, Pratt, and Price (2012).

explain why stores often set prices that end with 99 cents. Left-digit bias is typically evoked in settings where consumers are processing many numbers and are inattentive to all digits. In the marathon setting, we think it is unlikely that runners do not fully process their finishing time. However, it is possible that a runner's audience does not fully process the finishing times of others. A model of 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.). In the next section, we show 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 amount of bunching that occurs at 15-minute marks since there is no change in the left digit at those marks. We argue that reference points established at round numbers is a more natural psychological explanation our pattern of data.

Below we test whether round numbers actually act as reference points.

4 Results

4.1 Excess mass at round numbers

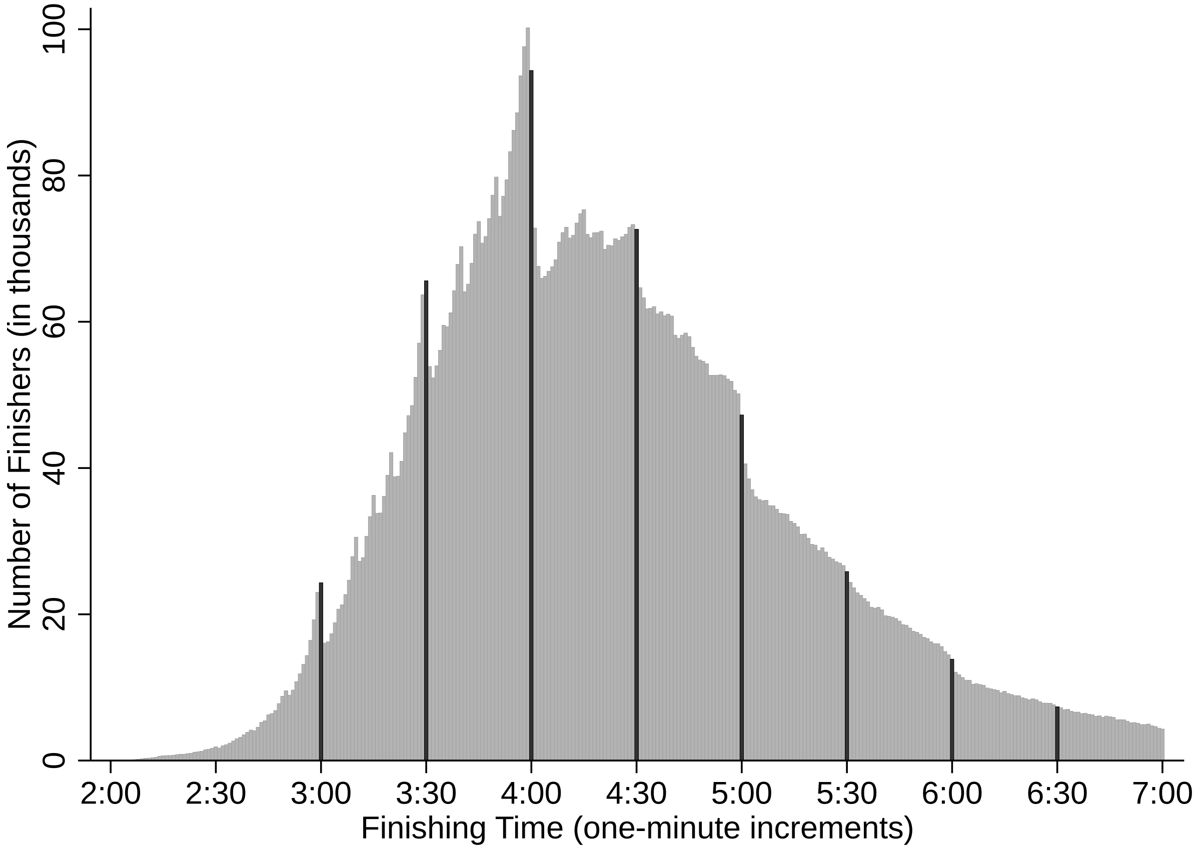
Figure 2 provides the distribution of finishing times (with one-minute bins) for our full sample of runners. The highlighted bars are the bins in the minute just prior to every 30-minute mark (e.g., 3:59:00 to 3:59:59) and indicate clear excess mass just to the left of the 30-minute marks. For example, there are 96,099, 98,405, and 92,970 finishers in the minute before the 3:58, 3:59, and 4:00 marks, compared to 71,682, 66,630, and 65,049 finishers in the 4:00, 4:01, and 4:02 bins. While the 4-hour mark is particularly dramatic, qualitatively similar differences exist at other hour and half-hour marks, and to a lesser extent at 10-minute marks. There are 51.4%, 21.7%, and 29.6%, 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: 12.1%, 8.6%, 9.6%, and 7.2% for 3:10, 3:20, 3:40, and 3:50, respectively.[10]

We also fit a 15-order polynomial to the density function from 2:20 to 7:00, using 1 minute bins. The plot of residuals between the actual and fitted density functions in Figure 3 shows a similar pattern as Figure 2. The pattern is similar for different size bins and higher and lower order polynomial fits.

We use two approaches to formally test whether excess mass exists at round numbers in the distribution. The two approaches serve different purposes. First, we test whether there is a significant discontinuity in the density function at round numbers, and whether the largest discontinuity occurs at the round number, or

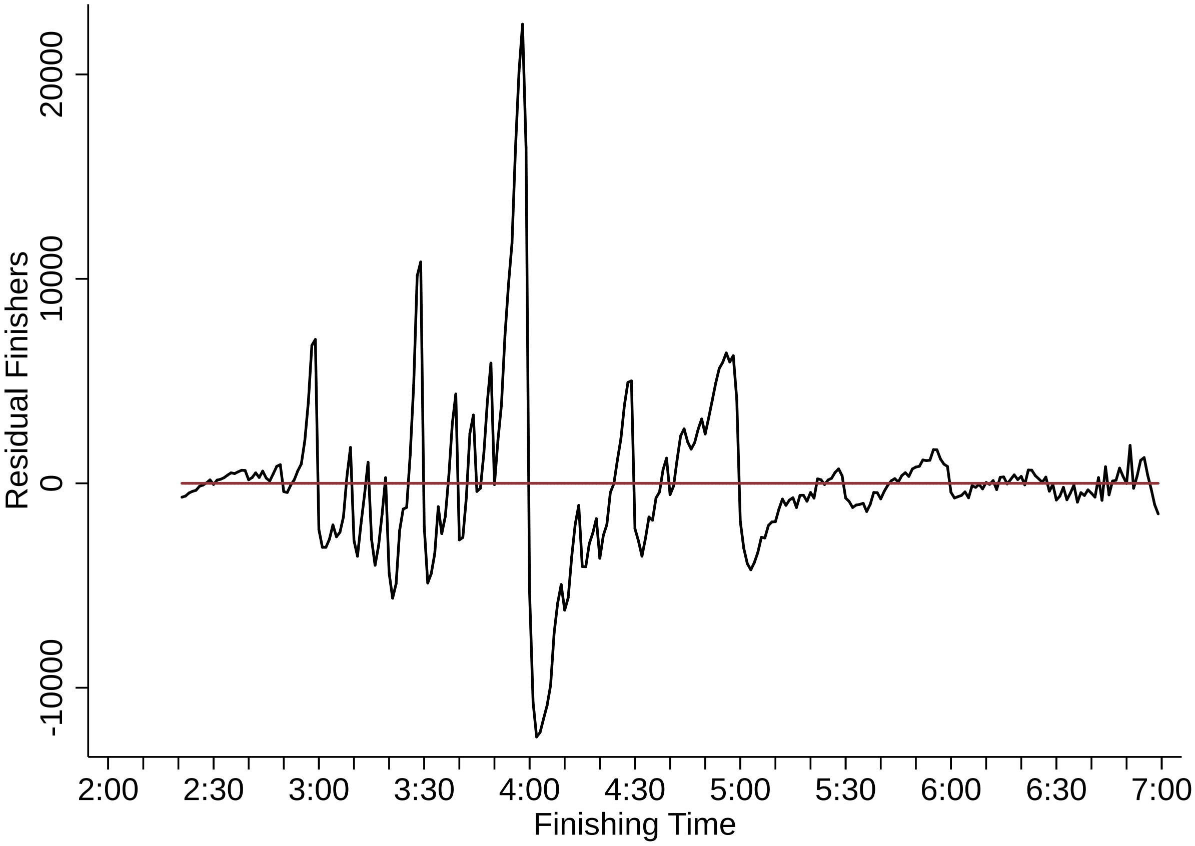
[10] Note that all of these thresholds are to the left of the median of the distribution. This excess mass measure would be negative for normal, lognormal and many other single-peaked continuous distributions.

Figure 2: Distribution of marathon finishing times ($n = 9,524,071$)



NOTE: The dark bars highlight the density in the minute bin just prior to each 30 minute threshold.

Figure 3: Residual finishers between actual and polynomial fit to finishing time



NOTE: A 15-order polynomial is fit to the density function of finishing times, discretized in 1 minute bins. For example, For example, there were 22,456 more actual finishers between 3:58:00 and 3:58:59 than indicated by the polynomial fit.

merely around the round number. Second, we measure the amount of excess mass around the round number reference point and test whether the excess mass is statistically significant.

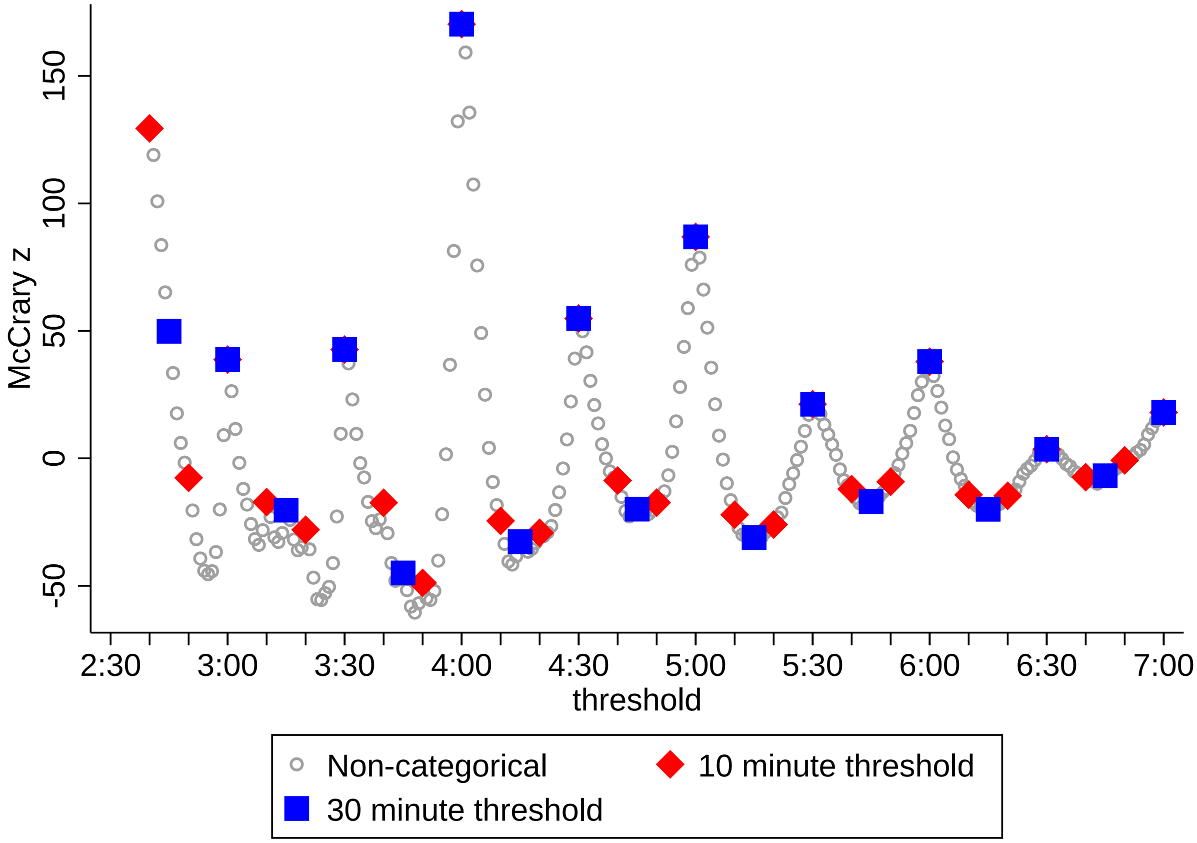
The first approach uses a method proposed by McCrary (2008) to test whether there is a discontinuity in the density function at some designated threshold. The regression discontinuity method estimates a locally linear density function on both sides of the threshold and uses bootstrap methods to determine whether that gap is statistically significant. Whereas the histogram of finishing times provides clear visual support that excess mass exists, the McCrary test provides a statistically precise manner of evaluating that claim.

In Figure 4, we plot a running z-statistic from the McCrary test, where we test for a significant jump in the density function at each minute mark between 2:40 and 7:00. The figure shows very large z-statistics at hour marks and half-hour marks. Note that the running z-statistic "jumps" at 3:00, 3:30, 4:00, etc., with each of the z-statistics at these round numbers a local optima. Although it is harder to detect visually from Figure 4, many of the 10-minute marks are local maximum, even though some of these z-statistics are negative. The negative values reflect the zero-sum nature of the McCrary statistic across the support of the density function. For example, the large z-statistic at 4 hours also implies significantly negative z-statistics once we move away from 4 hours in each direction. Thus, tests at one threshold are not independent of tests at neighboring thresholds.[11]

Although the McCrary test is useful for establishing a significant discontinuity in the density function at a reference point, we need another methodology to quantify how many runners are being displaced. To do this, we adopt the methodology proposed in Chetty, Friedman, Olsen, and Pistaferri (2011) to quantify the extent of excess mass in an interval around a round number. 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). That literature hypothesized that income will bunch around tax rate thresholds. Consistent with that hypothesis, Chetty, Friedman, Olsen, and Pistaferri (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 as 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 unlikely to be able 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

[11] To discriminate whether there are significant discontinuities at the 10-minute marks, we also conduct statistical tests which use just the density immediately around the round number. Specifically, we test whether the density in the Δt minutes to the left of the round number is significantly different from the sum of the density in the $\Delta t/2$ minute bins to the left and right of that interval. These tests show significant discontinuities at many 10-minute marks for Δt between 1 and 5 minutes. For example, for $\Delta t = 2$, all 10 minute marks between 2:30 and 6:00, except for 5:20 and 5:50, yield statistically significant differences at the conventional $p < .05$ level.

Figure 4: Running McCrary z-statistic



NOTE: The McCrary test is run at each minute threshold from 2:40 to 7:00 to test whether there is a significant discontinuity in the density function at that threshold.

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, we expect to see 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.

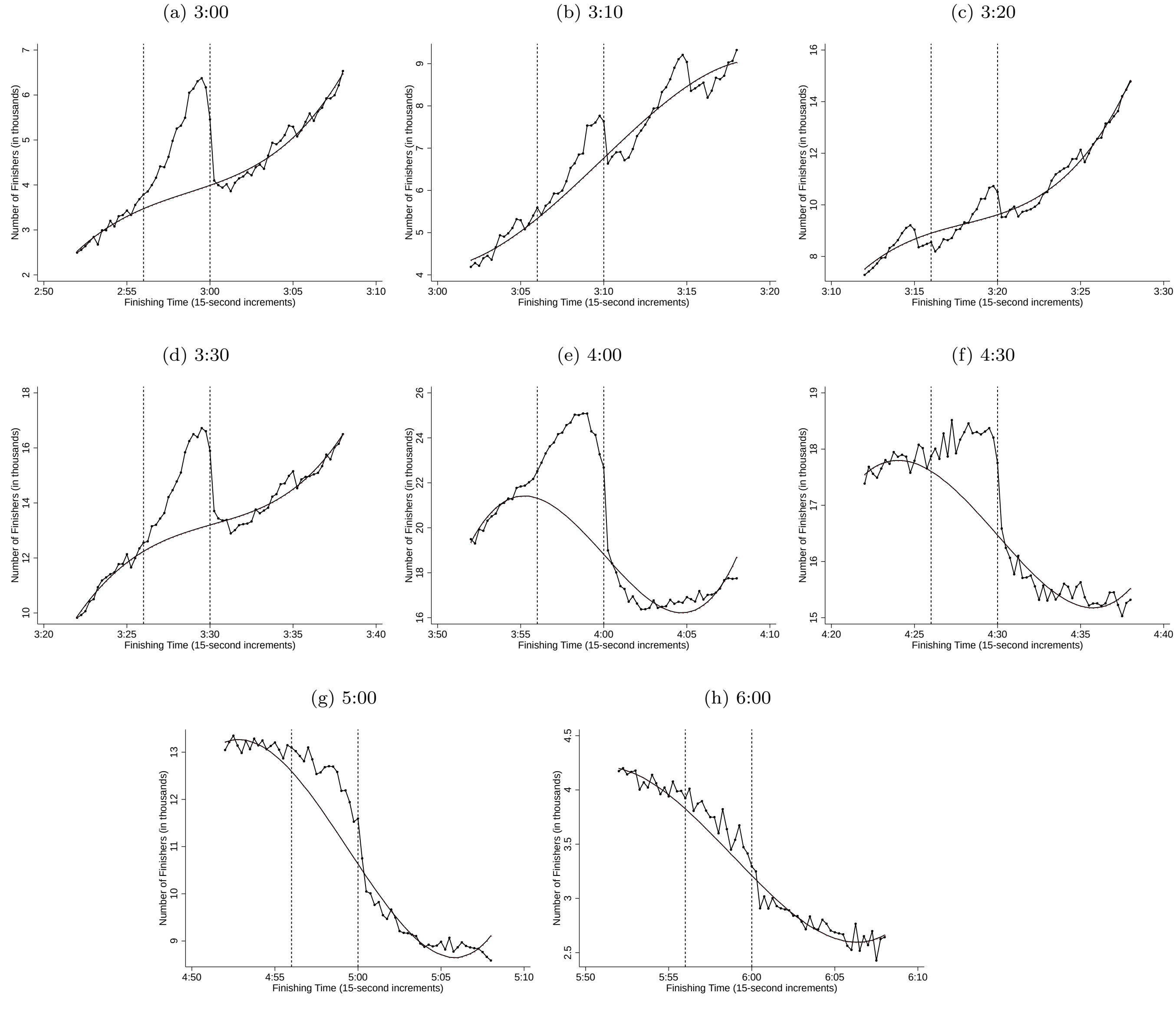
To calculate the amount of bunching, we follow the Chetty et al. (2011) methodology. The counterfactual distribution is estimated by fitting a cubic polynomial to the local density of finishing times around the reference point *excluding* the bunching region. The difference between the actual density 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. For our purposes, we consider 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, we look at a window from 3 hours and 22 minutes to 3 hours and 38 minutes in order to test for bunching at the 3 hours and 30 minute mark. The primary reason for choosing this window is that it avoids 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. This window was chosen based on visual inspection of the bunching (as recommended by Chetty et al. (2011)). 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 of 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 true counterfactual world.

The main results of the bunching estimation applied to our full sample are depicted in Figure 5 and summarized in Table 2. Figure 5 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.[12,13]

[12] Our results are robust to variations in the window around a reference point as well as the bunching region for excess mass.

[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 and found considerably stronger results for chip time. To do so, we restricted our sample to runners with both clock and chip times ($n = 5,581,034$). Using the Chetty et al. (2011) procedure on this sample, we estimated 5.7% excess mass for clock time and 13.4% excess mass for chip time at 4 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 ($n = 4,032,505$) (2.9% excess

Figure 5: Distribution of the number of finishers around round number reference point and the fitted counterfactual distribution


(a) 3:00

(b) 3:10

(c) 3:20

(d) 3:30

(e) 4:00

(f) 4:30

(g) 5:00

(h) 6:00

NOTE: The vertical axis shows the number of finishers in each 15 second bin. The jagged line reflects the actual density function, while the smooth curve is the counterfactual density fitted using the Chetty et al. (2011) procedure. The "bunching region" starts 4 minutes before a round number and ends at the round number.

Table 2 provides summary measures from the procedure that is shown graphically in Figure 5. 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 500 iterations. The largest number of runners (47,363) is displaced into the bunching region at 4 hours, while the largest percentage increase in finishers (24.6%) occurs at 3 hours. We find statistically significant evidence of bunching for all 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.

It is important to note that our instantiation of the Chetty et al. 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 5 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 24.6% excess mass ($t = 41.2$) at 3 hours using a bunching region of [2:56,3:00]. The excess mass is 33.2% ($t = 40.6$) if we restrict the bunching region to [2:59,3:00] and 28.8% ($t = 52.9$) if the bunching region is [2:57,3:00].

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 3, 4, and 5 hours 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.

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 candidate for an extrinsic benefit at a round number is qualifying for the Boston Marathon. The annual Boston Marathon is one of the oldest marathons in the world and one of the

mass for clock time and 13.1% excess mass for chip time) or at least two minutes slower than their chip time ($n = 3,224,469$) (1.0% excess mass for clock time and 12.6% excess mass for chip time). We find similar results at other round numbers.

few major marathons that has a qualifying time (although runners may also participate by working through charitable organizations). In order to qualify to participate in the Boston Marathon, a runner must run a different marathon in a certain qualifying time, which is determined by a runner's age and gender. For example, from 2003 to 2012, the large majority of our sample, 18-34 year-old males had to run a marathon in under 3 hours and 10 minutes in order to qualify for the Boston Marathon. The qualifying time for females of the same age category 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.[14]

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 males between the ages of 60 and 64 and females between the ages of 45 and 49. Thus, 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 those whose age and gender indicate that the round number is not associated with a Boston Marathon qualifying time.[15] The last 2 columns in Table 1 indicate that these sample exclusions do very little to our estimates of excess finishers. The largest change is at the 3 hour and 10 minute mark where excess bunching drops from 7.2% to 5.0%.

4.3 Pace setters and peer effects

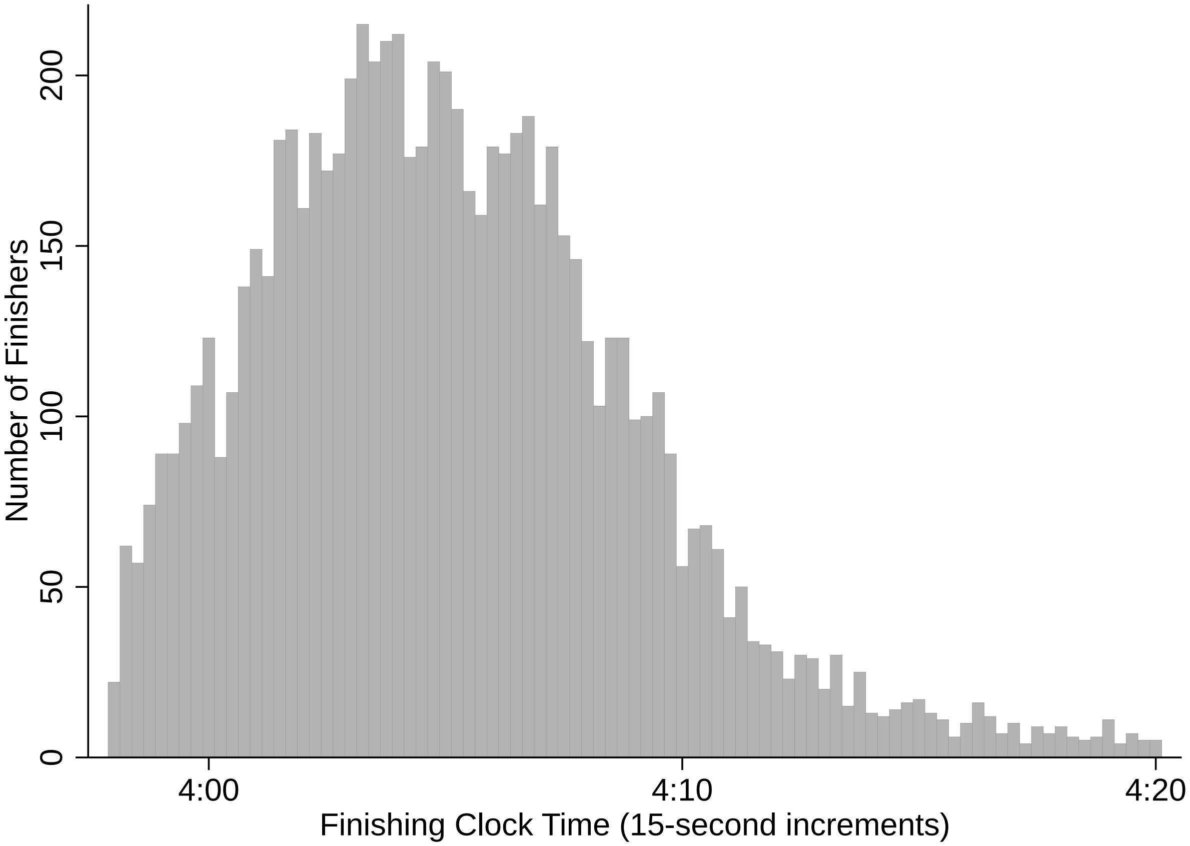
Most major marathons have pace teams. For example, the 2013 Chicago Marathon provided pace teams for 3:00, 3:05, 3:10, 3:15, 3:20, 3:25, 3:30, 3:35, 3:40, 3:45, 3:50, 3:55, 4:00, 4:10, 4:25, 4:30, 4:40, 4:55, 5:00, 5:10, 5:25 and 5:45. If having a pace setter is valuable, then the institutional feature of pace teams could be an alternative explanation for the bunching we observe at round numbers. While pacing groups is a reasonable alternative hypothesis, several pieces of evidence suggest that this cannot be the major driver of the effects that we find. For example, the results we present in the next subsection on late race effort provision are difficult to explain with pace setters. However, additional evidence suggest that this cannot be an explanation for our results.

In large marathons such as the Chicago Marathon, runners cross the starting line at very different clock

[14] From 1997 to 2012, the Boston Marathon rounded times down, and thus a time of 3:10:59 qualified that 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] (6.2%) than for [3:07,3:11] (4.1%).

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

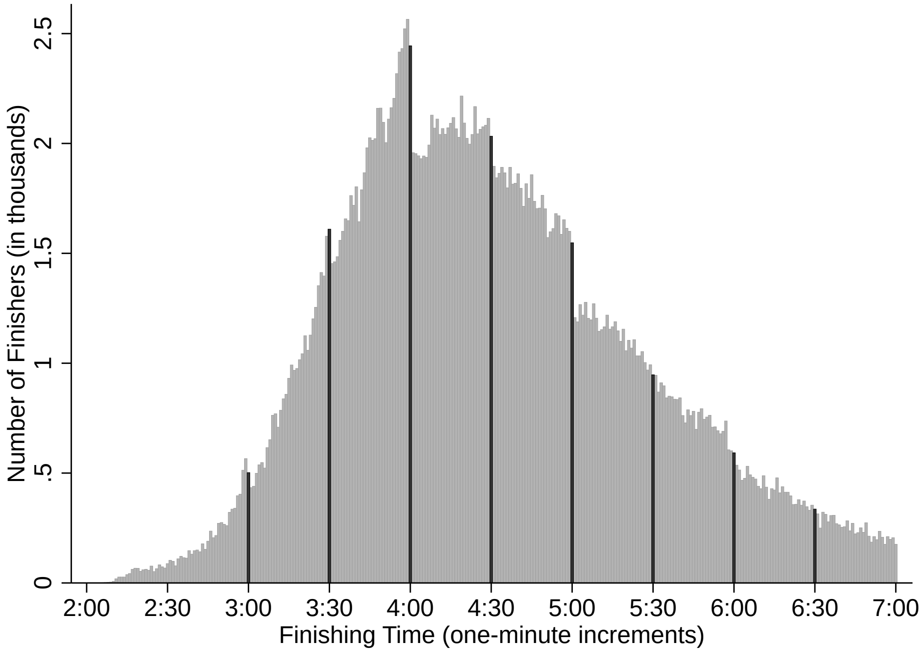
Figure 6: Clock time for runners with a chip time between 3:58 and 4:00, Chicago Marathon, 1998-2011



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 the bunching that we find, then we should see a large group of runners who cross the finish line at the same clock time (since pace setters can only work if you are physically in the same area as them). In Figure 6, 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. (Clock times are similarly spread out if we plot the distribution for each year.) The fact that the runners who are contributing to the bunching just before the 4-hour mark 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 pace setters as the primary driver of our results is to focus on marathons that almost surely do not have institutionalized pace teams. We do so by focusing on small marathons. Figure 7 plots the distribution of finishing times for marathoners ($n = 291,231$) who participated in one

Figure 7: Histogram of finishing times for marathons with less than 200 finishers



NOTE: The dark bars highlight the density in the minute bin just prior to each 30 minute threshold.

of the 3,268 marathons with fewer than 200 finishers.[16] 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.[17] Table 3 shows that bunching at the hour marks is substantial for marathons held prior to 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 suggest that peer effects cannot be driving the bunching results, since the average difference in finishing times between one runner and the next runner is 2.58 minutes for all runners and 1.46 minutes for runners finishing between 3 hours and 50 minutes and 4 hours and 10 minutes.

[16] Indeed, the website `findmymarathon.com` indicates that none of these marathons have pace teams.

[17] `http://runcim.org/got-pacers-you-bet/`

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.

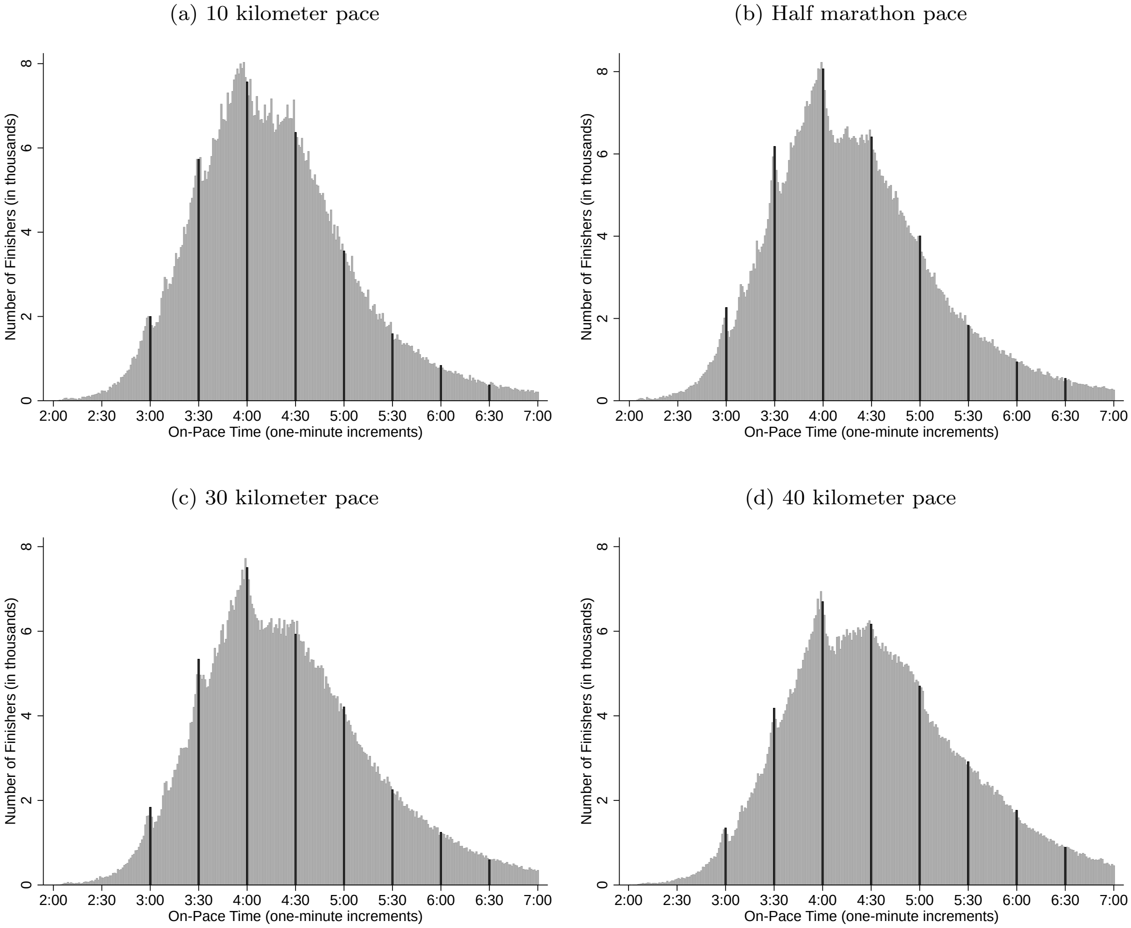
Similarly, a marathon runner who has reference-dependent preferences can employ a number of different strategies that each could create bunching at the reference point. We use the richness of the marathon data to examine effort provision throughout various stages of the race. In particular, we explore 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 performance toward the end of the race in order 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 unround 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 8 shows the distribution of finishing times linearly extrapolated from paces 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 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 the runners advance further in the marathon. For example, the excess mass percentages and t-statistics from the Chetty et al. analysis at the 3-hour marks are: 5.1% and 4.42 (10 kilometers), 13.7% and 9.42 (half marathon), 12.4% and 7.74 (30 kilometers), 14.1% and 6.38 (40 kilometers), and 28.2% and 16.26 (finish). This analysis indicates that at least part of the final bunching that we find is due to runners 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 full marathon pace for the first 40 kilometers. We term this measure a runner's normalized pace. A runner's full marathon pace is calculated by multiplying their 40 kilometer split by 42.195/40. Panel A of Figure 9 plots the relative pace against the 40 kilometer pace, focusing on the runners who are on pace to finish in around 4 hours. The vertical axis in Panel A indicates that runners

Figure 8: Histogram of extrapolated finishing times, based on intermediate splits


(a) 10 kilometer pace

(b) Half marathon pace

(c) 30 kilometer pace

(d) 40 kilometer pace

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

ran the final 2.195 kilometers of a marathon 6-11% 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, and 4:05 and 4:15 ran approximately 8-11% 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 6-8% 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 9 zooms out to show this normalized pace for runners across the entire 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 at a quicker pace when they are close to a round number than when they are further away.

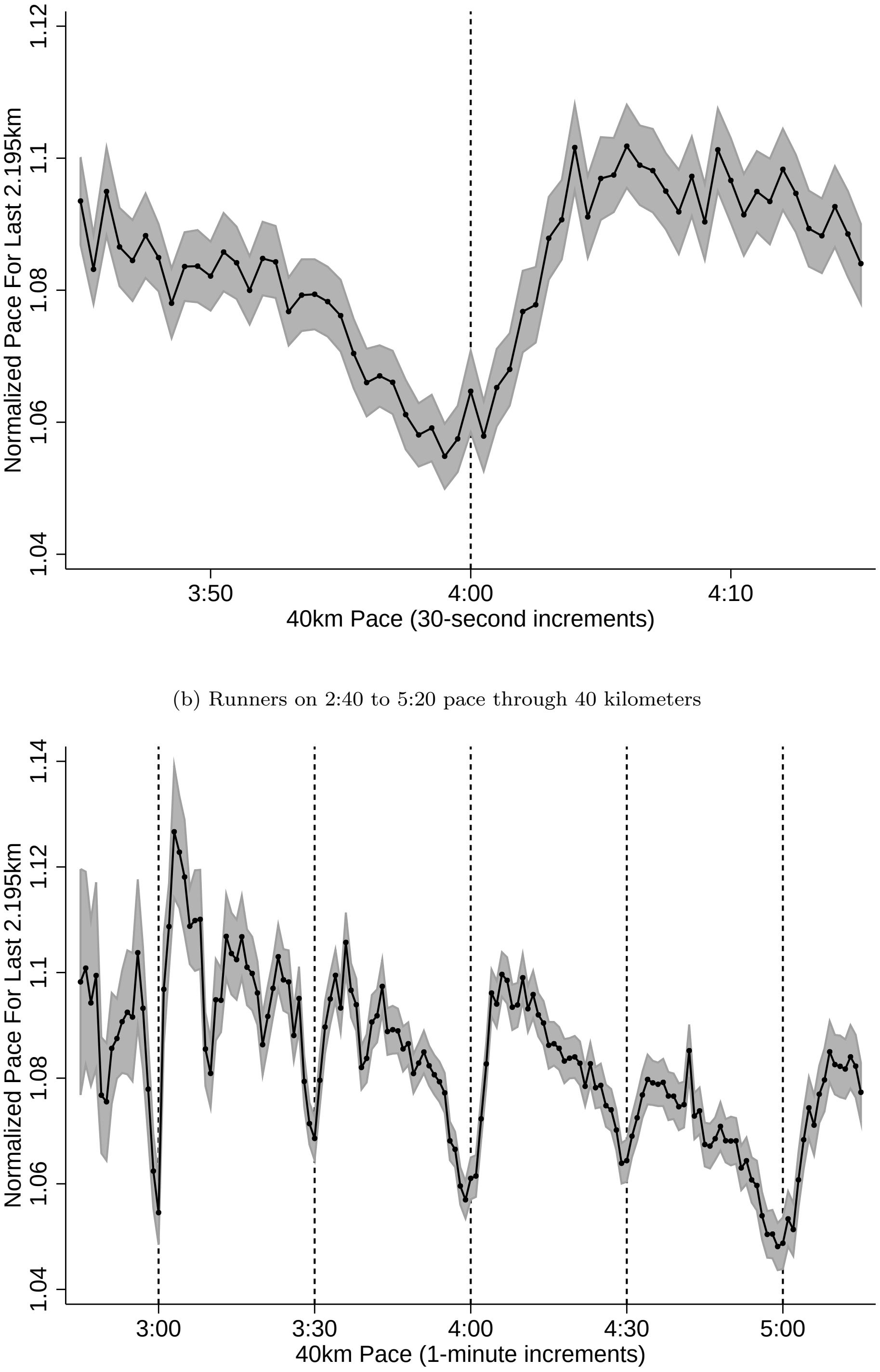
Note that this pattern of effort allocation is not due to runners choosing different strategies for allocating energy. For example, the pattern could be explained by a mixture of runners who run quicker from 30 to 40 kilometers, expending more effect so they can coast in, and runners who are more conservative in order to save up 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. To the contrary, the correlation between pace from 30 to 40 kilometers and pace over the last 2.195 kilometers is .63. The correlation remains large ($\rho = .49$) and positive even when we drop the 25% of runners who slow down the most from 30 to 40 kilometers. Indeed, the pattern documented in Figure 9 holds if we normalize the last 2.195 kilometer pace relative to the pace from 30 to 40 kilometers.

Figure 9 shows that effort provision in the last 2.195 kilometers of a race depends heavily on a runner's proximity to a round-numbered 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 10 and 11.

Figure 10 plots the probability that a runner runs the last 2.195 kilometers at a faster pace than the first 40 kilometers. Panel A indicates that about 25% of runners increase their speed in the last 2.195 kilometers. This fraction, however, increases to approximately 35% if the runner was right on target to finish at a round number. Once again, Panel B shows the likelihood of speeding up across the entire 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 7.9% slower over the last 2.195 kilometers. In Figure 11, we show the probability that a runner

Figure 9: Normalized pace for last 2.195 kilometers as a function of 40 kilometer pace

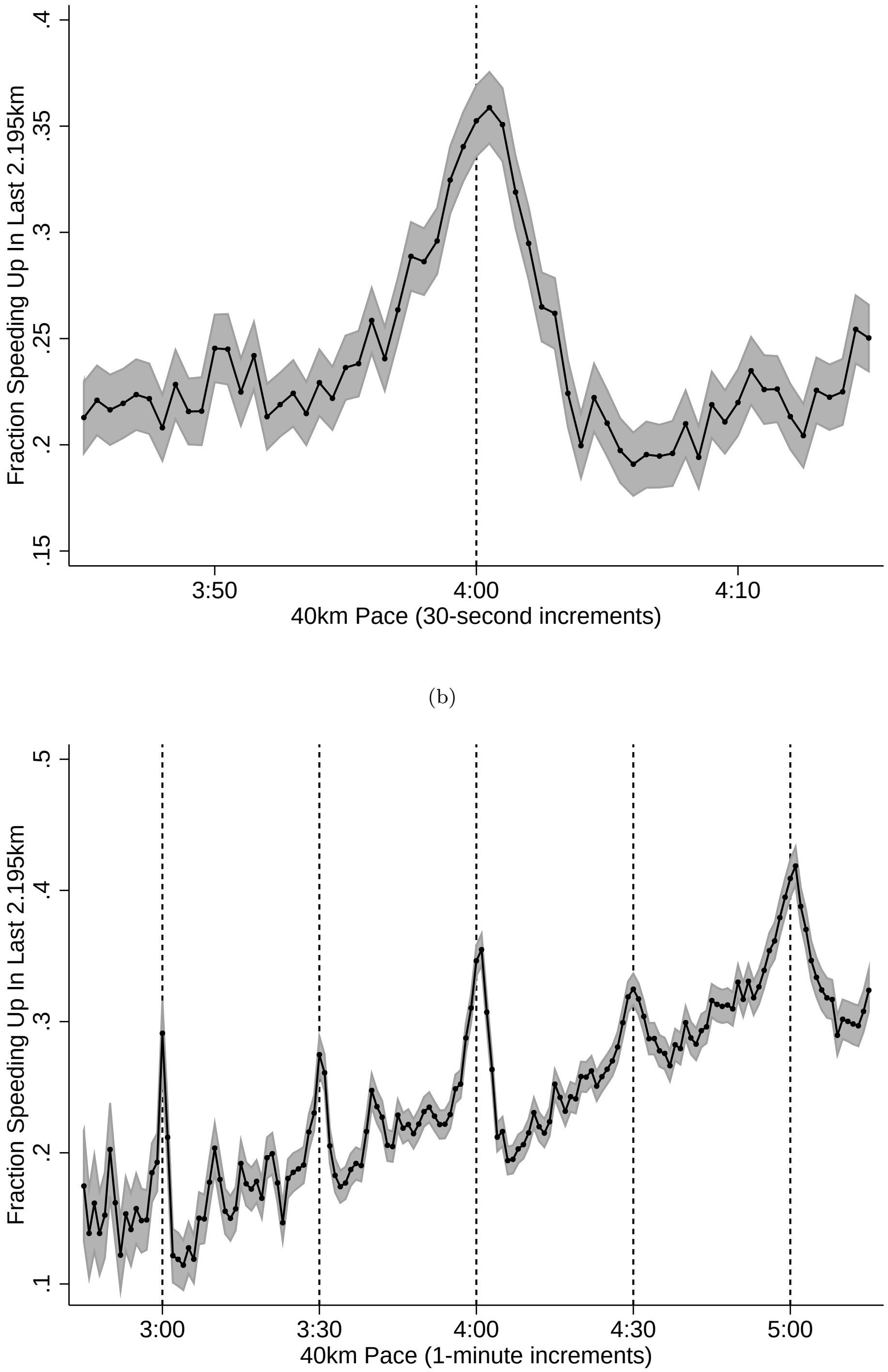


(a) Runners on 3:45 to 4:15 pace through 40 kilometers

(b) Runners on 2:40 to 5:20 pace through 40 kilometers

NOTE: 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.

Figure 10: Percentage of marathoners who speed up over the last 2.195 kilometers



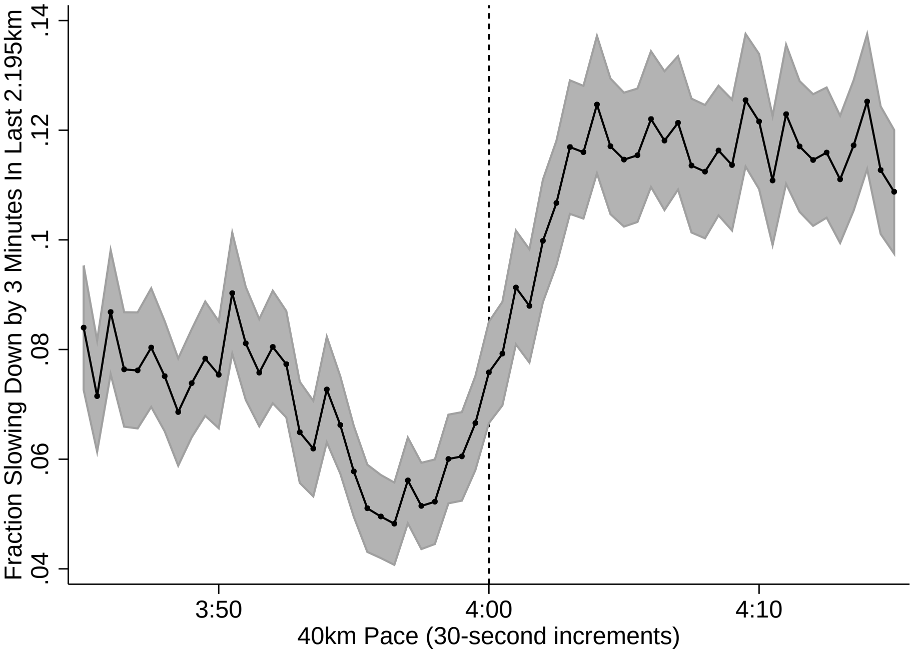
(a)

(b)

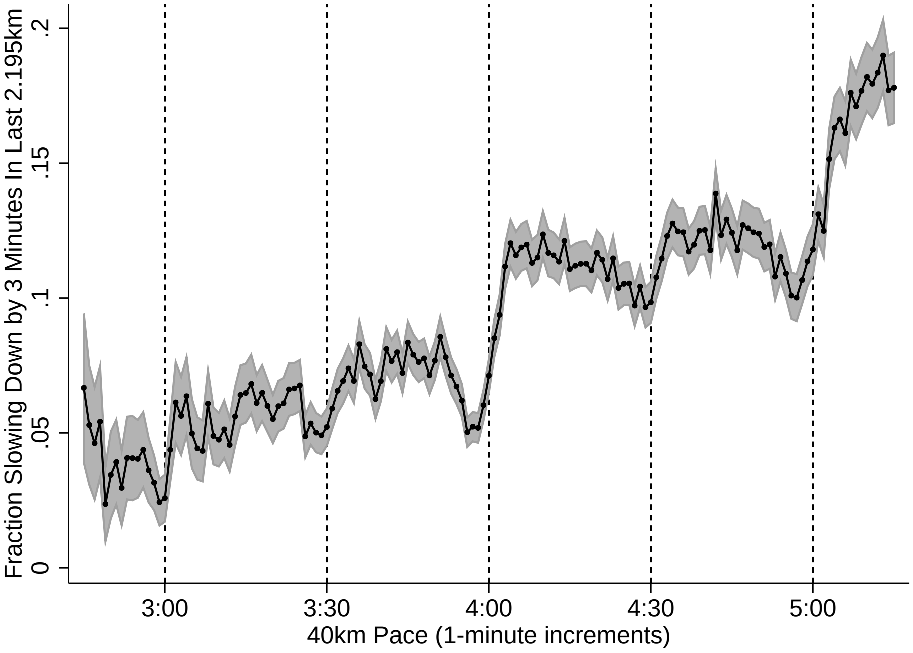
NOTE: 95% confidence intervals are depicted by the shaded regions.

Figure 11: Percentage of marathoners who slow down over the last 2.195 kilometers by 3 or minutes in full marathon pace

(a)



(b)



NOTE: 95% confidence intervals are depicted by the shaded regions. A runner slows down by 3 minutes or more if their full marathon pace through 40 kilometers is t and their full marathon pace for the final 2.195 kilometers is $t + 3$ or more.

slowed his or her pace in the last 2.195 kilometers by more than 3 minutes. Panel A depicts this probability near the 4-hour mark, while Panel B looks at the entire 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 leg of the marathon than runners who were not in range to finish ahead of the reference point.

Collectively, Figures 8 to 11 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.

4.5 Spontaneous Goal Formation

The results that we have presented so far suggest that runners have round-number reference points and exert effort in order to finish just better than these reference points. We showed that this can occur through pacing and planning as well as dynamic effort provision at the end of the race. All of this evidence is consistent with runners establishing a reference point prior to running the race. For example, a runner may choose a 4-hour goal and either pace to meet that mark, or try to stay reasonably close to a 4-hour pace and then exert effort at the end of the race to surpass the reference point.

Is it possible that as the race progresses, runners adopt reference points that differ from those established at the start of the race? If so, this is likely to occur when a runner is no longer able to beat their initial reference point and therefore may spontaneously adopt a new round number target.

We test whether runners spontaneously adopt new reference points during the race by focusing on runners who fall substantially behind their early pace. We then test whether these runners still are more likely to finish just before a slower round number.

Figure 12 shows graphs based on the Chetty et al. (2011) method that are analogous to those in Figure 5. Figure 12, however, restricts the sample to runners who were on pace to run at least 15 minutes faster than the reference point shown in that figure through the half marathon. For example, panel A shows the histogram of finishing times around 3 hours and 30 minutes. This histogram is restricted to runners who were on pace to finish faster than 3 hours and 15 minutes at the half-marathon mark. These figures provide evidence of a small but noticeable bunching for a set of runners who slow considerably. The t-statistics for the 3:30, 4:00, 4:30, and 5:00 marks are 2.52, 6.14, 5.02, and 2.32 respectively.

Although most runners do not set out to achieve a goal by running much faster in the first half and then much slower in the second half, this pacing may still reflect the strategy of some runners. To deal with this concern, we look at a factor which leads almost all runners to run slower, extremely hot weather. The 2007 Chicago Marathon was extraordinarily hot, with the temperature approaching 90 degrees Fahrenheit

(32 degrees Celsius) during the race. Between 1998 and 2013, excluding 2007, 81.0% runners ran the second half of the Chicago Marathon slower than the first half. In 2007, 98.7% of runners slowed down. In 2007, runners ran 28.9% slower in the second half of the marathon than in the first. For all other years, the ratio was 10.7%. In addition, the 2,410 runners who finished in both 2006 and 2007 ran 41.87 minutes (or 17.6%) slower in 2007 than 2006.

Thus, presumably many, if not most of the runners in 2007, abandoned their original goals. Indeed, only 3.2% of the 2007 Chicago Marathon runners in Sackett, Wu, White, and Markle's (2014) sample met their goal. On average, runners finished 51.46 minutes short of their goal, compared to 14.15 minutes for the remaining 14 marathons in that study. We examine whether the 2007 Chicago Marathon weather shock nevertheless led runners to bunch at round-number reference points. Figure 13 plots the running McCrary statistic. Although the pattern is not as stark as in Figure 4, we still see statistically signficant "jumps" in the McCrary statistic at 4:00, 4:30, and 5:00.

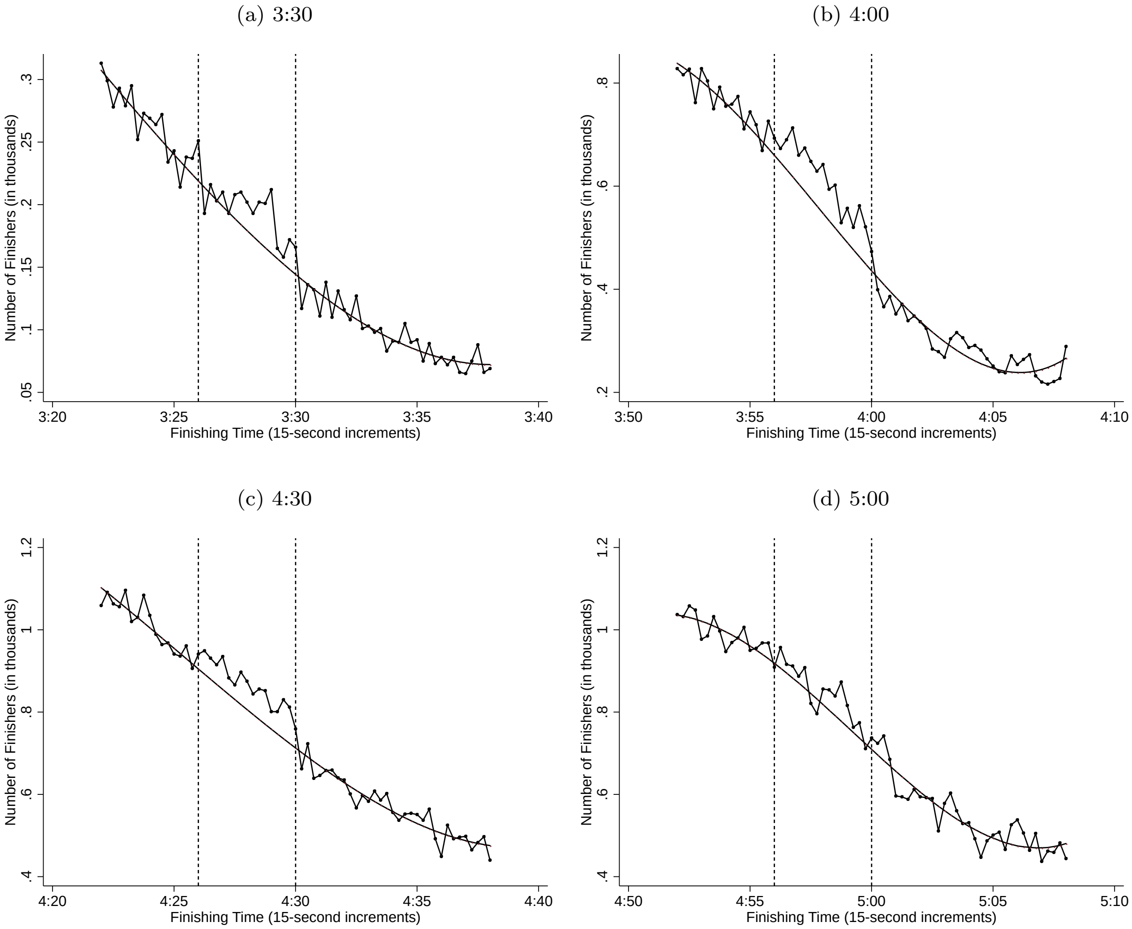
In sum, the analysis of runners who slowed down and 2007 Chicago Marathon finishers suggest that runners adopt reference points spontaneously. Kőszegi and Rabin (2006) proposed that reference points are based on "beliefs... held in the recent past about outcomes." Our results on this section suggest that reference points, whether goals or expectations, can be adapted very quickly as other reference points recede from the set of possible outcomes.

5 Calibration Results

In this section, we perform a calibration exercise to examine whether the observed amount of bunching and excess mass is consistent with standard prospect theory parameters. In Section 2, we proposed that three forms of reference dependence can lead to bunching of finishing times near a reference point. In this exercise, we primarily examine the second and third discontinuities, kinks in the first and second derivative of the benefit function. We do so because we can reference the large literature on measurement of the prospect theory value function (e.g., Abdellaoui, Bleichrodt, and Paraschiv, 2007; Gonzalez and Wu, 1999; Tversky and Kahneman, 1992). At the end of this section, however, we also calibrate a benefit function with a jump at the reference point.

To calibrate our levels of excess mass, we use a functional form commonly employed in this literature for

Figure 12: Spontaneous Goal Formation Histograms


(a) 3:30

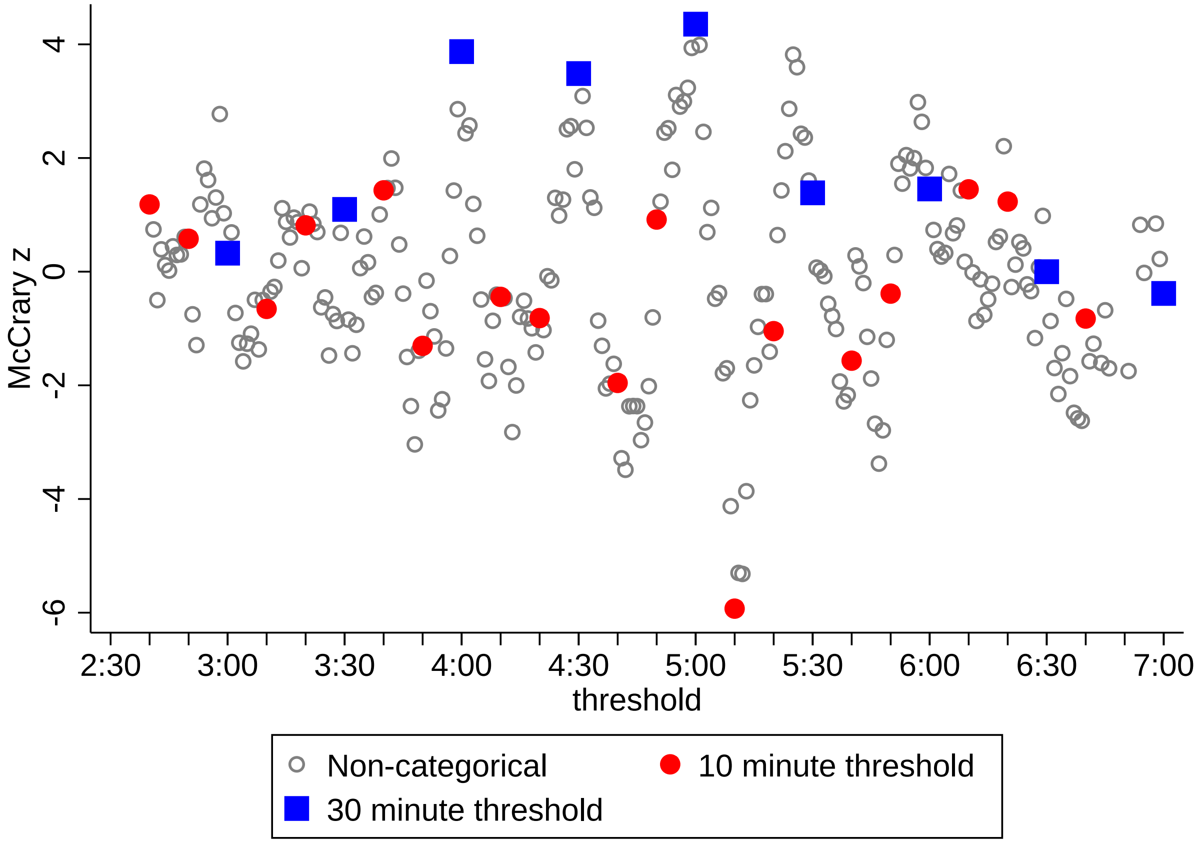
(b) 4:00

(c) 4:30

(d) 5:00

NOTE: Sample is restricted to runners on pace to run 15 minutes or more faster than 3:30, 4:00, 4:30 and 5:00. For example, in Panel (A), all runners were on pace to run 3:45 or faster through the half marathon. The histograms follow the same method as used in Figure 5.

Figure 13: Running McCrary z-statistic for 2007 Chicago Marathon



NOTE: The McCrary test is run at each minute threshold from 2:40 to 7:00 to test whether there is a significant discontinuity in the density function at that threshold.

estimating the prospect theory value function,

$$v(x) = \begin{cases} x^{\alpha_G}, & x > 0 \\ -\lambda(-x)^{\alpha_L}, & \text{otherwise} \end{cases},$$

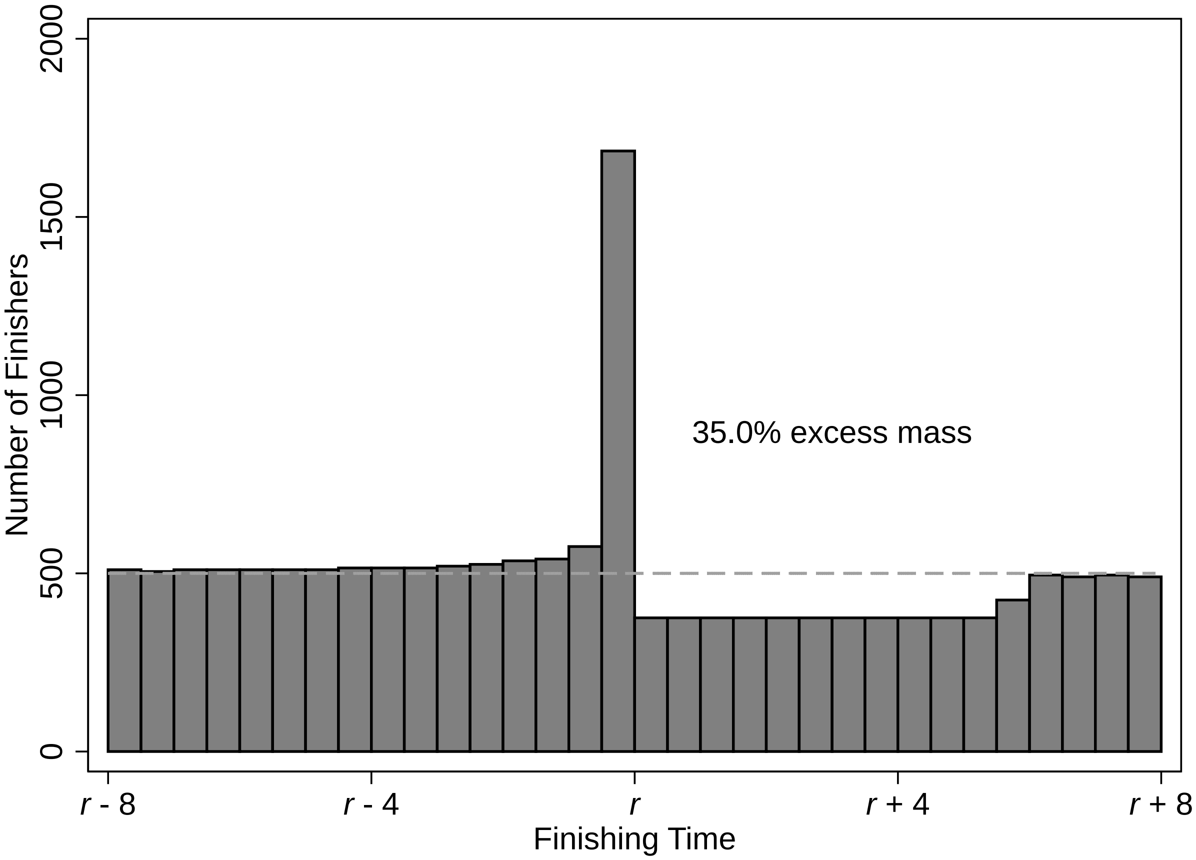
where 0 is the reference point. In this parametric specification, $\lambda > 1$ indicates loss aversion, and $\alpha_G, \alpha_L < 1$ indicates diminishing sensitivity. Tversky and Kahneman (1992) used certainty equivalent data for risky gambles and estimated $\hat{\lambda} = 2.25$ and $\hat{\alpha_G} = \hat{\alpha_L} = .88$. Other researchers using different procedures and datasets have found similar estimates (e.g., Abdellaoui et al., 2007). In our calibration exercise, we make the simplifying assumption that $\alpha_G = \alpha_L$, which Tversky and Kahneman showed is a good approximation (see, however, Nilsson, Rieskamp, and Wagenmakers, 2011).

We use this functional form and examine how excess mass is related to λ, $\alpha \doteq \alpha_G = \alpha_L$, as well as the proportion of agents who have reference-dependent preferences. As in Section 2, we let τ be a runner's finishing time, and $t = k - \tau$ be the number of minutes that a runner finishes ahead of k minutes, the worst possible finishing time. We set $k = 600$ and $r = 240$ in our calibration exercise.

To conduct our calibration exercise, we need to scale the marginal benefit and cost function. We assume that the marginal cost function has the following form, $c'(t) = \omega t^{\beta}$, where ω reflects differences in abilities across individuals and β is held constant across runners. The exponent, β, captures how sensitive finishing times are to changes in marginal benefits, with low β indicating more sensitivity. We assume that $\beta = 100$, which with loss aversion of $\lambda = 2.25$ increases motivation sufficiently to improve a finishing time from 4:03:05 to 3:59:59. Because this scaling is somewhat arbitrary, we examine the robustness of our results by varying β.

We assume that some proportion p of runners have reference-dependent preferences, with $1-p$ of runners having "classical preferences." For simplicity, we take classical preferences to be a linear benefit function, $b(t) = t$, or $b'(t) = 1$. We then back out the distribution of ω values that yields a uniform distribution of finishing times on $[r - 8, r + 8]$ in the absence of reference dependence. We allow reference-dependent preferences to redistribute finishing times from a wider interval, $[r-16, r+16]$, because diminishing sensitivity pulls finishing times toward the reference point and there might otherwise be a discontinuity near $r - 8$ or $r + 8$ as a result. In accord with Proposition 2.3, we scale the benefit function by choosing μ such that $\mu 8^{\alpha} = 8$, so that $b(r + 8) - b(r)$ is constant for all α. In contrast to Section 4, we know the counterfactual distribution and therefore can easily compute the excess mass on any interval. We also measure excess mass on $[r - 4, r]$, as we did for the Chetty et al. analysis in Section 4.

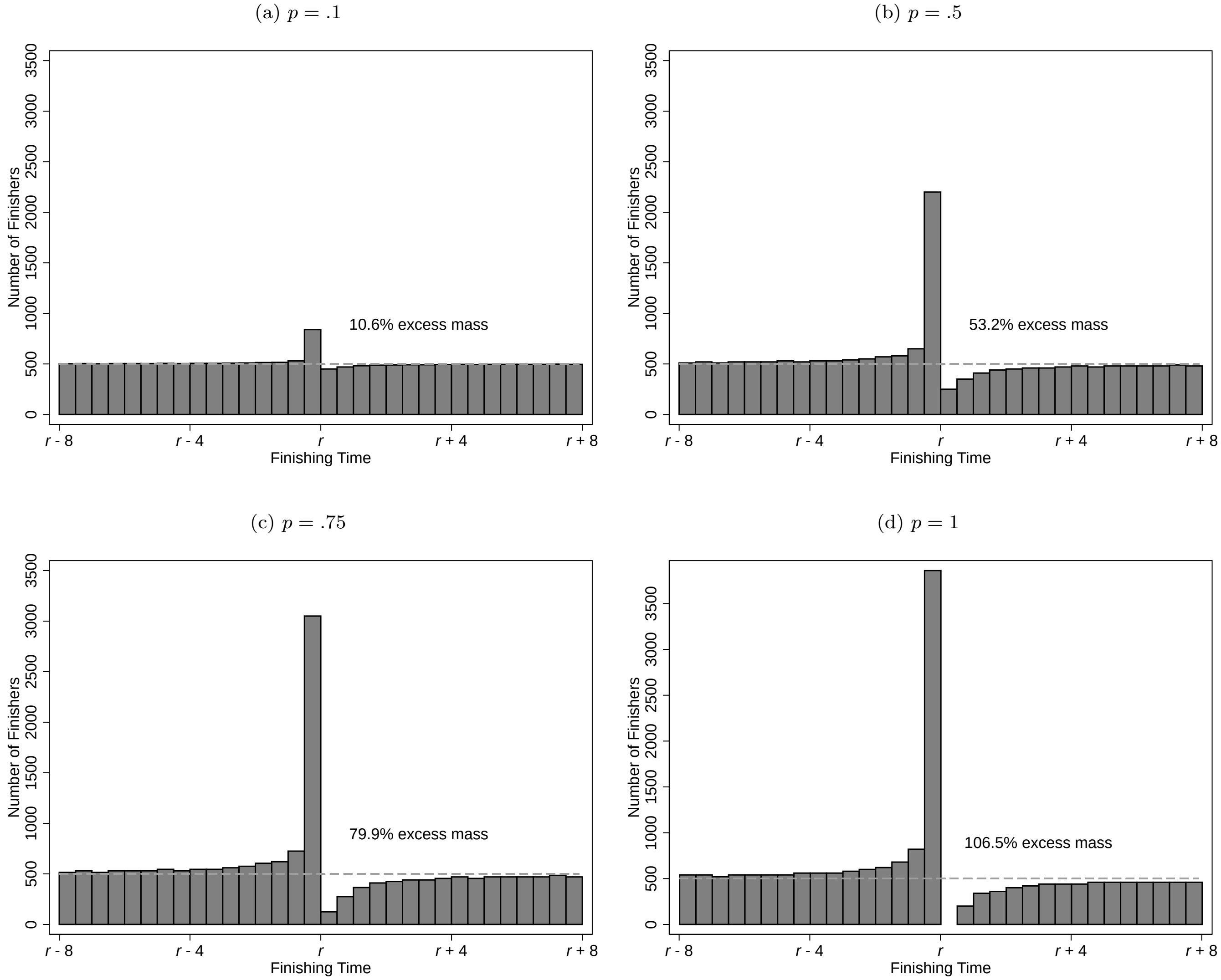
Figure 14: Calibration of finishing times, assuming $\lambda = 2.25$, $\alpha = .88$, and 25% reference-dependent runners



NOTE: The dotted line indicates the counterfactual distribution, with $p = 0$. Our calibration exercise assumes that classical preferences yield a uniform distribution with 500 finishers for each 30 second bin. Excess mass is measured from $r - 4$ to r.

For a given loss aversion parameter, λ, and diminishing sensitivity parameter, α, we can determine each runner's optimal performance numerically. Figure 14 shows the distribution of finishing times for $\lambda = 2.25$, $\alpha = .88$, and $p = .25$. We can see that these parameter values reproduce the basic qualitative shape of the distribution shown in the panels of Figure 5. The major differences are that the distribution bunches just to the left of the reference point, r, and the amount of excess mass is somewhat larger than estimated by the Chetty et al. procedure, even for 3 hours (see Table 2). Both of these differences reflect in part the deterministic nature of our calibration exercise, which assumes that runners maximize according to their deterministic benefit and cost functions. As mentioned earlier, actual runners face a stochastic control problem. Building in uncertainty about the marginal cost function and risk aversion would make Figure 14 look more like Figure 5. Below, we show that lower values of α also produce a diffuse density to the left of the reference point.

Figure 15: Calibration results, varying the proportion of runners with reference-dependent preferences



(a) $p = .1$

(b) $p = .5$

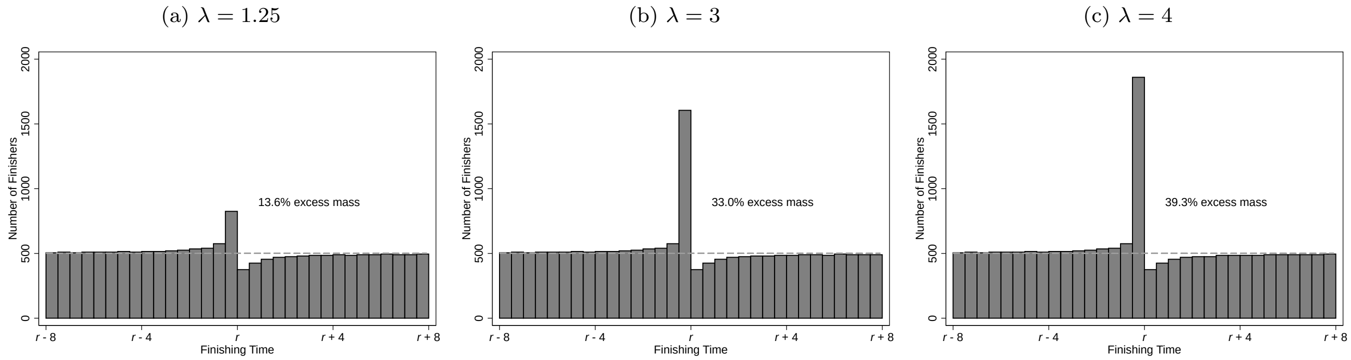
(c) $p = .75$

(d) $p = 1$

We also examine how this distribution of finishing times changes as we vary the proportion of runners with reference-dependent preferences (Figure 15), plotting the distribution for p = .1, .5, .75, and 1. As expected, excess mass increases in p. However, there is significant excess mass, even when the proportion of reference dependent runners is relatively low ($p = .1$). In all cases, most of the bunching is just to the left of the reference point, r.

We also vary λ (Figure 16), α (Figure 17), and β (Figure 18). Quite intuitively, excess mass is monotonically increasing in λ. As α decreases, consistent with Proposition 2.3, density shifts toward r from above and below r. Thus, Figure 17, Panel (a) indicates that while risk and uncertainty can produce diffuse bunching, so can diminishing sensitivity. Finally, we see that β changes the amount of excess mass, with low β showing

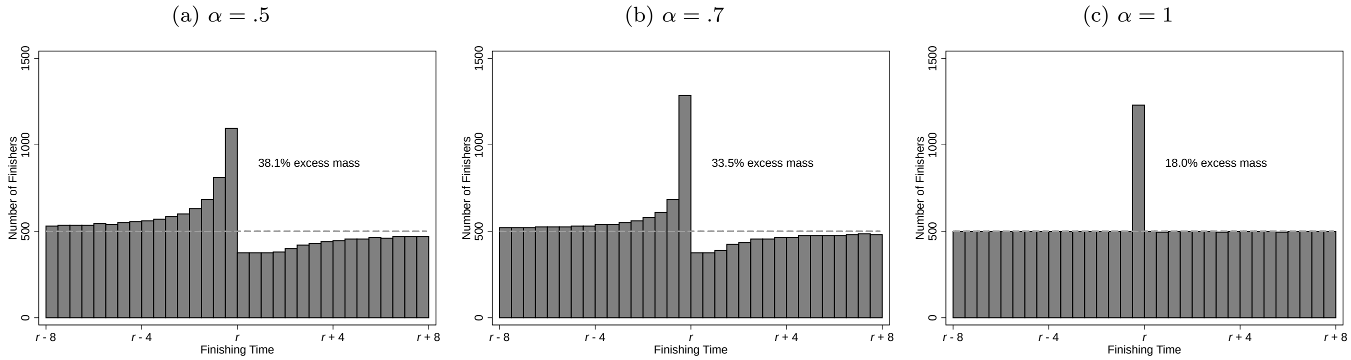
Figure 16: Calibration results, varying the loss aversion parameter, λ


(a) $\lambda = 1.25$

(b) $\lambda = 3$

(c) $\lambda = 4$

Figure 17: Calibration results, varying the coefficient of diminishing sensitivity, α


(a) $\alpha = .5$

(b) $\alpha = .7$

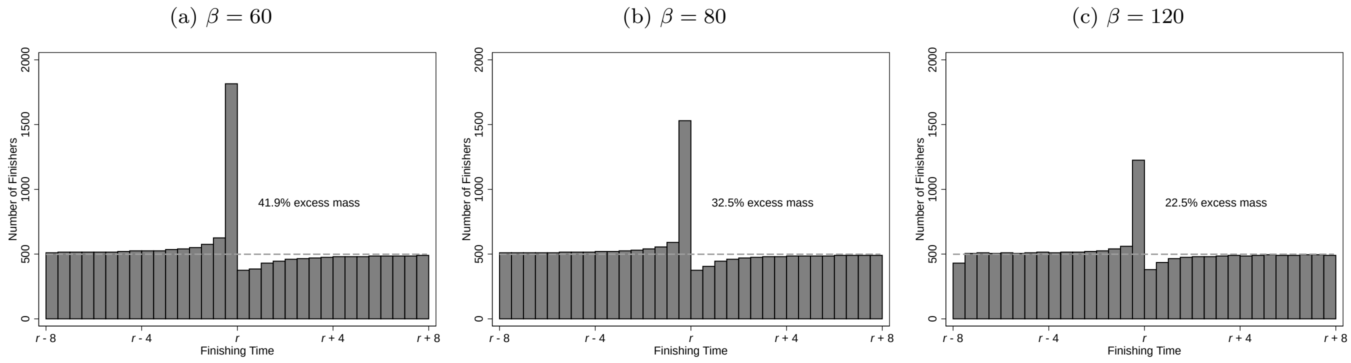
(c) $\alpha = 1$

the most sensitivity to changes in the marginal benefit function. We see that λ and β are essentially interchangeable in producing the same pattern of finishing times, as indicated by the virtually identical shapes between Figure 16, Panel (b) and Figure 18, Panel (b). To help visualize the relative role of λ, α, and p, we plot some iso-excess mass curves in Figure 19. These figures show how different combinations of parameters produce excess mass of 5%, 10%, 20%, 30%, and 40%.

We also repeat the same exercise, using the counterfactual densities shown in Figure 5 around 3 and 4 hours. We use a grid search to identify the λ, α, and p that best fit the actual density functon as determined by minimizing mean squared error. We set $\beta = 100$ throughout. The calibrated and actual densities are shown in Figure 20. For 3 hours, the best fitting parameters are: $\lambda = 2.35$, $\alpha = 0.23$, and $p = .12$. For 4 hours, the best fitting parameters are: $\lambda = 1.77$, $\alpha = 0.15$, and $p = .06$. Figure 20 shows that our calibration exercise is able to capture the basic contours of actual finishing times around 3 and 4 hours, albeit with lower α values than have been found in the literature.

We also note that a jump or notch in the benefit function as in Figure 1, Panel (a), produces a similar

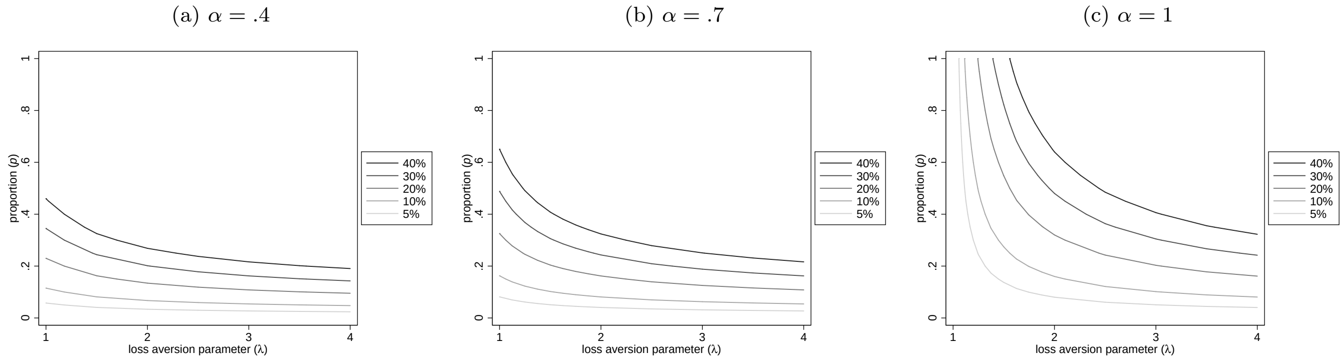
Figure 18: Calibration results, varying the coefficient, β, of the marginal cost function


(a) $\beta = 60$

(b) $\beta = 80$

(c) $\beta = 120$

Figure 19: Iso-excess mass curves showing combinations of α, λ, and p parameters that produce 5%, 10%, 20%, 30%, and 40% excess mass

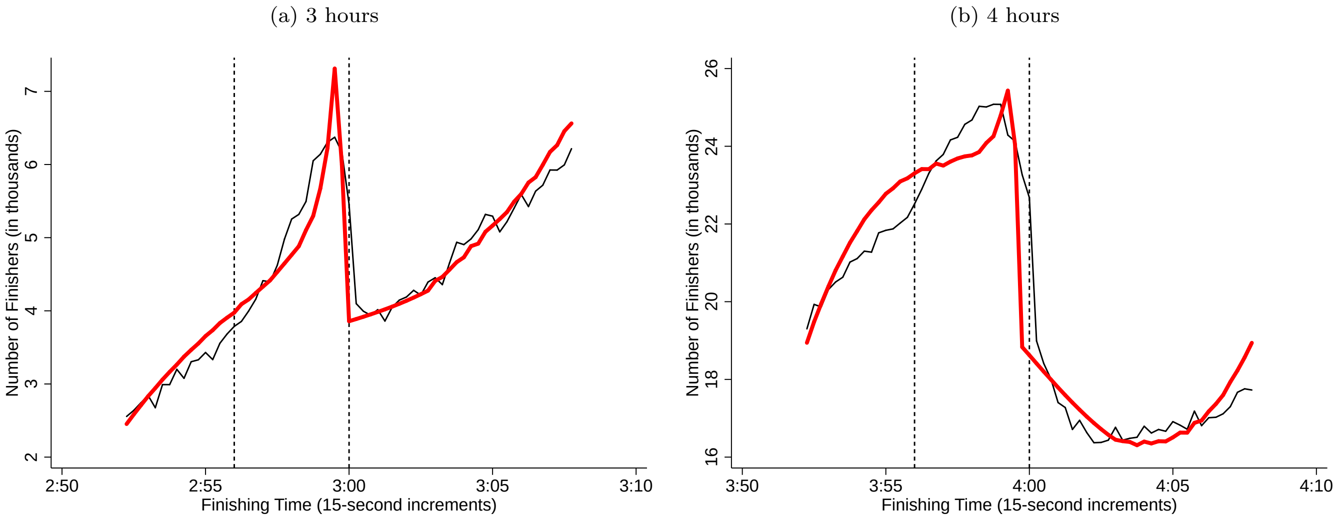

(a) $\alpha = .4$

(b) $\alpha = .7$

(c) $\alpha = 1$

NOTE: These calibrations set $\beta = 100$.

Figure 20: Calibration results fit to actual densities around the 3- and 4-hour thresholds



(a) 3 hours

(b) 4 hours

NOTE: The actual density is shown in black, with the thicker red line depicting the calibrated density function.

qualitative pattern as loss aversion. The calibration of such a jump in Figure 21 sets $p = .25$, $\alpha = .88$, and $\beta = 100$, and fixes the jump to produce an identical amount of excess mass as in Figure 14. There is one notable difference between Figures 14 and 21: a jump in the benefit function only improves performance for individuals who can surpass the reference point, whereas loss aversion increases the marginal benefit in losses and thus slides performance to the left for all runners.[18]

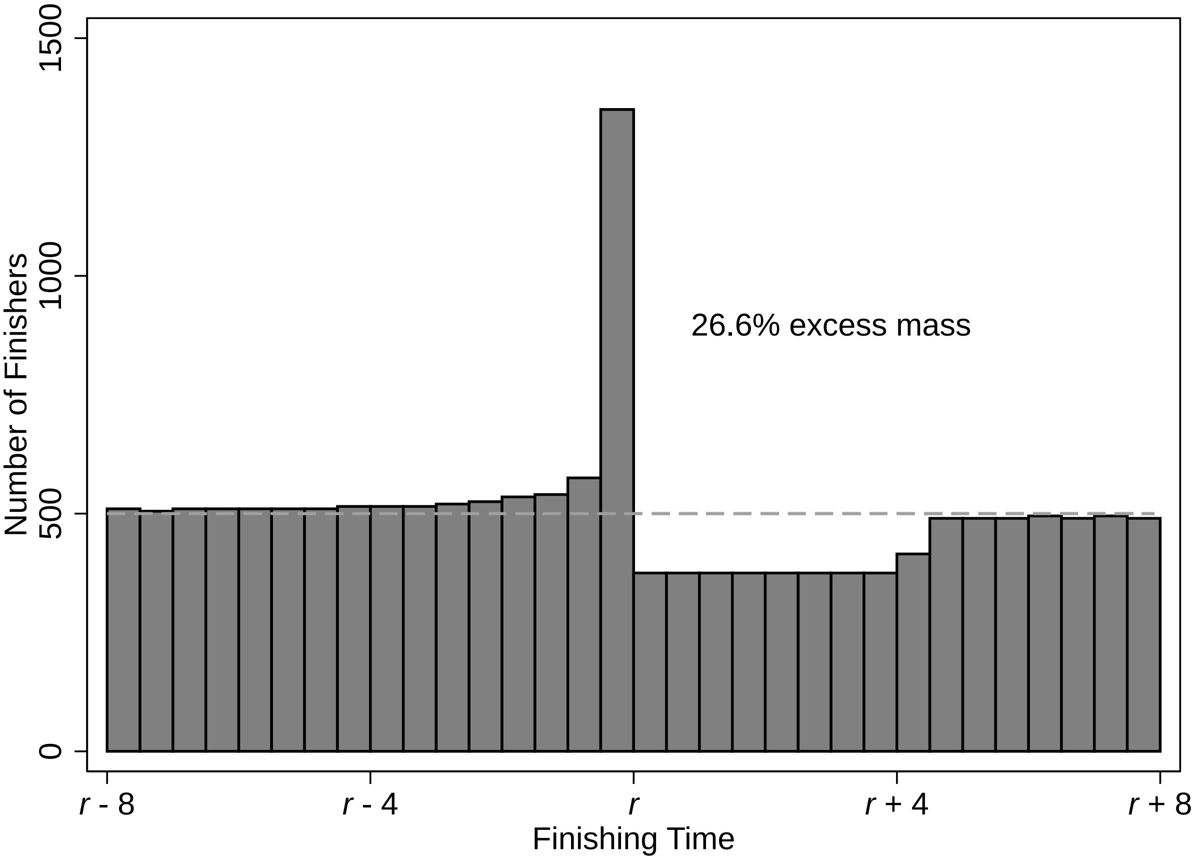
In sum, the calibration exercise indicates that relatively standard prospect theory parameters can reproduce the general pattern of bunching of finishing times discussed above. It is important to note that our calibration exercise is relatively silent about the proportion of runners who have classical and reference-dependent preferences, since λ, p, and β function similarly in our exercise. The exercise also demonstrates that we need somewhat more diminishing sensitivity than Tversky and Kahneman (1992) estimated to produce the diffuse pattern of bunching we observed in our archival data, although the diffuse pattern of finishing times just ahead of the reference point can also be captured by risk aversion.[19]

This calibration exercise also offers reasons why the degree of bunching varies across different round-number reference points. For example, why is the excess mass around 3 hours and 20 minutes so much

[18] Because we know the counterfactual distribution, the calibration exercise also allows us to examine whether the procedure used for estimating excess mass in Section 4.1 is biased. We calculated the actual excess mass, as well as the excess mass estimated from using the Chetty et al. procedure for 80 combinations of λ, α, and p and the prospect theory functional form. In all cases, the Chetty procedure underestimated the actual excess mass. On average, the actual excess mass was 28.0%, with the Chetty et al. procedure estimating 22.8%. We repeated the exercise for a jump in the benefit function. In this case, the Chetty et al. procedure slightly overestimated the excess mass (Actual: 34.8%; Chetty: 30.5%). Because there is likely to be a combination of a kink and a notch in the benefit function, the aggregate bias is probably small relative to the size of our effect.

[19] Although the α parameter used in Panels (a) and (b) of Figure 17 are lower than found in Tversky and Kahneman (1992), Gonzalez and Wu (1999) estimated $\hat{\alpha} = .49$.

Figure 21: Calibration of finishing times, assuming a jump in the utility function as in Figure 1, $\alpha = .88$, 25% reference-dependent runners, and $\beta = 100$.



NOTE: The dotted line indicates the counterfactual distribution, with $p = 0$. Our calibration exercise assumes that classical preferences yield a uniform distribution with 500 finishers for each 30 second bin. Excess mass is measured from $r - 4$ to r.

less than 3 hours (see Table 2)? The calibration provides some candidate explanations. For example, fewer runners may evoke 3 hours and 20 minutes as a reference point than 3 hours (i.e., $p_{3:20} < p_{3:00}$). Alternatively, 3:00 may be a "stronger" reference point, perhaps because 3 hours is "rounder" than 3 hours and 20 minutes, leading 3 hour runners to be more loss averse or have more diminishing sensitivity ($\lambda_{3:00} > \lambda_{3:20}$ or $\alpha_{3:00} < \alpha_{3:20}$). The iso-excess mass curves shown in Figure 19 provide parameter pairings that can produce the excess mass amounts that we find at different round number reference points.

6 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 three mechanisms, planning and pacing, reference-dependent effort provision near the finish line, and spontaneous goal formation. Although we do not report on these analyses here, we observe similar patterns for shorter racing distances, such as 10 miles and half marathon. 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.

Below we discuss some issues raised by these findings, as well as contributions of this research to the economics literature on reference dependence and self-control.

6.1 Aggregate effects of round number reference points

We have provided survey and behavioral evidence for the existence of round number reference points, such as 4 hours. However, we have been silent about: (i) whether the bunching draws "from above" or "from below"; and (ii) why marathoners, or more generally economic agents, adopt round number reference points. We start by addressing the first question, turning to the second question later.

The simple model presented in Section 2 demonstrates that reference dependence may increase performance or "draw from below" (Propositions 2.1 and 2.2) or decrease performance or "draw from above" (Proposition 2.3). Although we have been somewhat silent about the source of the bunching, our analysis of normalized pace over the last 2.195 kilometers suggests that both forms of drawing may be operating. Panels (a) of Figures 9-11 show a dramatic increase in effort provision for runners on the cusp of finishing faster than 4 hours. In the same figures, we also see that runners who are on pace to run 3:55 or faster run considerably slower, perhaps playing it safe so as not to jeopardize the prospect of beating 4 hours. Similar patterns at other round-number reference points suggest that these round numbers are drawing from below as well as above throughout the full range of performance.

We also fit parametric distributions and high-order polynomials to finishing times and examined the residuals relative to these parametric forms (e.g., Figure 3). These analyses suggest that the bunching caused by round number reference points draws from above and below. However, quantifying the relative proportion of each kind of displacement relative to the other is highly sensitive to parametric assumptions. Finally, our calibration results also suggest that some diminishing sensitivity in gains is probably needed to capture the diffuse bunching around round numbers, although we have also noted that this pattern could reflect risk attitudes instead.

6.2 Contributions to reference-dependence literature

Previous field studies of reference dependence have hypothesized that some well-defined quantity serves as a reference point (e.g., purchase price in Odean, 1998, or 52-week high stock price in Heath, Huddart, and Lang, 1999); manipulated reference points (e.g., Hossain and List, 2012; Fryer et al, 2012); or estimated a numeric reference point (e.g., Camerer et al, 1997; Crawford and Meng, 2011). Our study is most similar in this respect to the work on the disposition effect (Shefrin and Statman, 1985; Odean, 1998; Genosove and Mayer, 2001). We, like Shefrin and Statman (1985), hypothesized that some clear quantity (purchase price in their case and round numbers in our case) serves as a reference point, and looked for and found behavioral evidence consistent with this hypothesis. However, our setting extends previous research on reference dependence by proposing and examining a reference point that is not based on rational expectations (unlike Kőszegi and Rabin, 2006) and is endogenous to the economic agent (unlike Odean, 1998, or Genosove and Mayer, 2001).

6.3 Goal-setting in Economics

Goals are related but psychologically distinct from expectations and are often optimistic (Sackett et al, 2014). In this section, we discuss briefly the endogeneity of goals. A puzzle in the goal-setting literature is why individuals set optimistic goals, given that optimistic goals induce a generally disappointing comparison between performance and the reference point (Heath, Larrick, and Wu, 1999). Some recent theoretical work in economics has proposed a potential solution. This work builds on planner-doer models of the self akin to principal-agent models (e.g., Thaler and Shefrin, 1981). For example, recent models by Hsiaw (2013) and Koch and Nafziger (2011) propose that goals (set by "planners") can address the self-control or present-bias problems of agents (or “doers”). Although not a perfect analogy, goals may thus play a similar role in self-control problems as external rewards do in principal-agent problems.[20]

The proposal that goals are solutions to the self-control problem may also provide an explanation for why round numbers are adopted as reference points. In contrast to Pope and Simonsohn’s (2011) suggestion that round numbers serve as cognitive reference points (Rosch, 1975), we speculate that a coarse partition of the goal space (as in round numbers) may also be a defense against backsliding of present-based agents.

6.4 Other Margins and Other Reference Points

We have restricted our attention to the effect of reference points on the provision of effort and the distribution of performance. Of course, reference points likely have an impact on other margins. Indeed, we find suggestive evidence that a runner is significantly more likely to run in next year’s version of the same marathon if they run 4 hours and 1 minutes than if they run 3 hours and 59 minutes. However, a more comprehensive investigation of the effect of reference points on other extensive margins is beyond the scope of this paper.

In this setting, as in most natural field settings, other standards, besides round numbers, might also serve as reference points. We propose that there is 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 non-round 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 goals and hence reference points, and, of course, we know when a number is in fact round.

[20] See Brunnermeier and Parker, 2005, for a non-goal based account of why agents may have optimistic, and not rational, expectations.

Bibliography

Abdellaoui, Mohammed; Han Bleichrodt and Corina Paraschiv. 2007. "Loss Aversion under Prospect Theory: A Parameter-Free Measurement." *Management Science*, 53(10), 1659-74.

Anderson, Eric T. and Duncan Simester. 2003. "Effects of $9 Price Endings on Retail Sales: Evidence from Field Experiments." *Quantitative Marketing and Economics*, 1(1), 93-110.

Asch, Beth J. 1990. "Do Incentives Matter-the Case of Navy Recruiters." *Industrial and Labor Relations Review*, 43(3), 89S-106S.

Ashenfelter, Orley; Kirk Doran and Bruce Schaller. 2010. "A Shred of Credible Evidence on the Long-Run Elasticity of Labour Supply." *Economica*, 77(308), 637-50.

Austin, James T. and Jeffrey B. Vancouver. 1996. "Goal Constructs in Psychology: Structure, Process, and Content." *Psychological Bulletin*, 120(3), 338-75.

Barberis, Nicholas C. 2013. "Thirty Years of Prospect Theory in Economics: A Review and Assessment." *Journal of Economic Perspectives*, 27(1), 173-96.

Barberis, Nicholas and Richard Thaler. 2003. "A Survey of Behavioral Finance," George M. Constantinides, Milton Harris and Rene M. Stulz, *Handbook of the Economics of Finance*. Elsevier, 1053-128.

Barseghyan, Levon; Francesca Molinari; Ted O'Donoghue and Joshua C. Teitelbaum. 2013. "The Nature of Risk Preferences: Evidence from Insurance Choices." *American Economic Review*, 103(6), 2499-529.

Benartzi, Shlomo and Richard H. Thaler. 1995. "Myopic Loss-Aversion and the Equity Premium Puzzle." *Quarterly Journal of Economics*, 110(1), 75-92.

Berger, Jonah and Devin Pope. 2011. "Can Losing Lead to Winning?" *Management Science*, 57(5), 817-27.

Bertrand, Marianne; Jessica Pan and Emir Kamenica. 2013. "Gender Identity and Relative Income within Households." National Bureau of Economic Research Working Paper Series, No. 19023.

Brunnermeier, Markus K. and Jonathan A Parker. 2005. "Optimal Expectations." *American Economic Review*, 95(4), 1092-1118.

Camerer, Colin; Linda Babcock; George Loewenstein and Richard Thaler. 1997. "Labor Supply of New York City Cabdrivers: One Day at a Time." *Quarterly Journal of Economics*, 112(2), 407-41.

Card, David; Alexandre Mas; Enrico Moretti and Emmanuel Saez. 2012. "Inequality at Work: The Effect of Peer Salaries on Job Satisfaction." *American Economic Review*, 102(6), 2981-3003.

Chetty, Raj; John N. Friedman; Tore Olsen and Luigi Pistaferri. 2011. "Adjustment Costs, Firm Responses, and Micro Vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records." *Quarterly*

Journal of Economics, 126(2), 749-804.

Crawford, Vincent P. and Juanjuan Meng. 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), 1912-32.

DellaVigna, Stefano. 2009. "Psychology and Economics: Evidence from the Field." *Journal of Economic Literature*, 47(2), 315-72.

Diecidue, Enrico and Jeroen Van De Ven. 2008. "Aspiration Level, Probability of Success and Failure, and Expected Utility." *International Economic Review*, 49(2), 683-700.

Falk, Armin and Andrea Ichino. 2006. "Clean Evidence on Peer Effects." *Journal of Labor Economics*, 24(1), 39-57.

Farber, Henry S. 2005. "Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers." *Journal of Political Economy*, 113(1), 46-82.

Farber, Henry S. 2008. "Reference-Dependent Preferences and Labor Supply: The Case of New York City Taxi Drivers." *American Economic Review*, 98(3), 1069-82.

Fehr, Ernst and Lorenz Goette. 2007. "Do Workers Work More If Wages Are High? Evidence from a Randomized Field Experiment." *American Economic Review*, 97(1), 298-317.

Fishburn, Peter C. 1977. "Mean-Risk Analysis with Risk Associated with Below-Target Returns." *American Economic Review*, 67(2), 116-26.

Fryer, Roland G., Jr.; Steven D. Levitt; John List and Sally Sadoff. 2012. "Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment." National Bureau of Economic Research Working Paper Series, No. 18237.

Genesove, David and Christopher Mayer. 2001. "Loss Aversion and Seller Behavior: Evidence from the Housing Market." *Quarterly Journal of Economics*, 116, 1233-60.

Gonzalez, Richard and George Wu. 1999. "On the Shape of the Probability Weighting Function." *Cognitive Psychology*, 38(1), 129-66.

Hardie, Bruce G.S.; Eric J. Johnson and Peter S. Fader. 1993. "Modeling Loss Aversion and Reference Dependence Effects on Brand Choice." *Marketing Science*, 12(4), 378-94.

Heath, Chip; Steven Huddart and Mark Lang. 1999. "Psychological Factors and Stock Option Exercise." *Quarterly Journal of Economics*, 114(2), 601-27.

Heath, Chip; Richard P. Larrick and George Wu. 1999. "Goals as Reference Points." *Cognitive Psychology*, 38(1), 79-109.

Hossain, Tanjim and John A. List. 2012. "The Behavioralist Visits the Factory: Increasing Productivity

Using Simple Framing Manipulations." *Management Science*, 58(12), 2151-67.

Hsiaw, Alice. 2013. "Goal-Setting and Self-Control." *Journal of Economic Theory*, 148(2), 601-26.

Kahneman, Daniel. 1992. "Reference Points, Anchors, Norms, and Mixed Feelings." *Organizational Behavior and Human Decision Processes*, 51(2), 296-312.

Kahneman, Daniel; Jack L. Knetsch and Richard H. Thaler. 1990. "Experimental Tests of the Endowment Effect and the Coase Theorem." *Journal of Political Economy*, 98, 1325-48.

Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47(2), 263-91.

Kleven, Henrik J. and Mazhar Waseem. 2013. "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan." *Quarterly Journal of Economics*, 128(2), 669-723.

Kőbberling, Veronika and Peter P. Wakker. 2005. "An Index of Loss Aversion." *Journal of Economic Theory*, 122(1), 119-31.

Koch, Alexander K. and Julia Nafziger. 2011. "Self-Regulation through Goal Setting." *Scandinavian Journal of Economics*, 113(1), 212-27.

Kőszegi, Botond and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics*, 121(4), 1133-65.

Kőszegi, Botond and Matthew Rabin. 2007. "Reference-Dependent Risk Attitudes." *American Economic Review*, 97(4), 1047-73.

Kőszegi, Botond and Matthew Rabin. 2009. "Reference-Dependent Consumption Plans." *American Economic Review*, 99(3), 909-36.

Lacetera, Nicola; Devin G. Pope and Justin R. Sydnor. 2012. "Heuristic Thinking and Limited Attention in the Car Market." *American Economic Review*, 102(5), 2206-36.

Larkin, Ian. 2014. "The Cost of High-Powered Incentives: Employee Gaming in Enterprise Software Sales." *Journal of Labor Economics*, 32(2), 199-227.

Lefgren, Lars; Brennan Platt, and Joseph Price. 2012. "Sticking with What (Barely) Worked: A Test of Outcome Bias." Working paper.

Levitt, Steven D.; John A. List; Susanne Neckermann and Sally Sadoff. 2012. "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance," National Bureau of Economic Research Paper Series, No. 18165.

Lien, Jaimie W. 2013. "Deciding When to Quit: Reference-Dependence over Slot Machine Outcomes", Working Paper.

March, James and Zur Shapira. 1987. "Managerial Perspectives on Risk and Risk Taking." *Management Science*, 33(11), 1404-18.

Markle, Alex B.; George Wu; Rebecca J. White and Aaron M. Sackett. 2014. "Goals as Reference Points in Marathon Running: A Novel Test of Reference-Dependence." Working paper.

Mas, Alexandre. 2006. "Pay, Reference Points, and Police Performance." *Quarterly Journal of Economics*, 121(3), 783-821.

Mas, Alexandre and Enrico Moretti. 2009. "Peers at Work." *American Economic Review*, 99(1), 112-145.

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

McGraw, A. Peter; Jeff T. Larsen; Daniel Kahneman and David Schkade. 2010. "Comparing Gains and Losses." *Psychological Science*, 21(10), 1438-45.

Moskowitz, Tobias and L. Jon Wertheim. 2011. *Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won*. Random House.

Murphy, Kevin J. 2000. "Performance Standards in Incentive Contracts." *Journal of Accounting and Economics*, 30(3), 245-78.

Odean, Terrance. 1998. "Are Investors Reluctant to Realize Their Losses?" *Journal of Finance*, 53(5), 1775-98.

Oyer, Paul. 1998. "Fiscal Year Ends and Nonlinear Incentive Contracts: The Effect on Business Seasonality." *Quarterly Journal of Economics*, 113(1), 149-85.

Pope, Devin G. and Maurice E. Schweitzer. 2011. "Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes." *American Economic Review*, 101(1), 129-57.

Pope, Devin and Uri Simonsohn. 2011. "Round Numbers as Goals." *Psychological Science*, 22(1), 71-79.

Pope, Devin G. and Justin R. Sydnor. forthcoming. "Behavioral Economics: Economics as a Psychological Discipline" in *The Blackwell Handbook of Judgment and Decision Making* (Gideon Keren and George Wu, editors).

Post, Thierry; Martijn J. van den Assem; Guido Baltussen and Richard H. Thaler. 2008. "Deal or No Deal? Decision Making under Risk in a Large-Payoff Game Show." *American Economic Review*, 98(1), 38-71.

Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature*, 37(1), 7-63.

Rees-Jones, Alex. 2013. "Loss Aversion Motivates Tax Sheltering: Evidence from U.S. Tax Returns." Working Paper.

Nilsson, Hkan; Jrg Rieskamp and Eric-Jan Wagenmakers. 2011. "Hierarchical Bayesian Parameter Estimation for Cumulative Prospect Theory." *Journal of Mathematical Psychology*, 55(1), 84-93.

Romer, David. 2006. "Do Firms Maximize? Evidence from Professional Football." *Journal of Political Economy*, 114(2), 340-65.

Rosch, Eleanor. 1975. "Cognitive Reference Points." *Cognitive Psychology*, 7(4), 532-47.

Sackett, Aaron M.; George Wu; Rebecca J. White and Alex B. Markle. 2014. "Harnessing Optimism: How Eliciting Goals Improves Performance." Working paper.

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

Samuelson, William and Richard Zeckhauser. 1988. "Status Quo Bias in Decision Making." *Journal of Risk and Uncertainty*, 1(1), 7-59.

Shefrin, Hersh and Meir Statman. 1985. "The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence." *Journal of Finance*, 40(3), 777-90.

Sydnor, Justin. 2010. "(Over) Insuring Modest Risks." *American Economic Journal: Applied Economics*, 2(4), 177-99.

Thaler, Richard H. and Hersh M. Shefrin. 1981. "An Economic Theory of Self-Control." *Journal of Political Economy*, 89(2), 392-410.

Tom, Sabrina M.; Craig R. Fox; Christopher Trepel and Russell A. Poldrack. 2007. "The Neural Basis of Loss Aversion in Decision-Making under Risk." *Science*, 315(5811), 515-18.

Tovar, Patricia. 2009. "The Effects of Loss Aversion on Trade Policy: Theory and Evidence." *Journal of International Economics*, 78(1), 154-67.

Triplett, Norman. 1898. "The Dynamogenic Factors in Pacemaking and Competition." *The American Journal of Psychology*, 9(4), 507-33.

Tversky, Amos and Daniel Kahneman. 1991. "Loss Aversion in Riskless Choice: A Reference Dependent Model." *Quarterly Journal of Economics*, 106(4), 1039-61.

Tversky, Amos and Daniel Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty*, 5(4), 297-323.

Wakker, Peter P. and Amos Tversky. 1993. "An Axiomatization of Cumulative Prospect Theory." *Journal of Risk and Uncertainty*, 7(2), 147-76.

Table 1: Summary statistics for full sample and full split sample

	Total Marathon Sample			Full Splits Sample		
	Mean	Std. Dev.	Observations	Mean	Std. Dev.	Observations
Finishing time (minutes)	4:27:00	0:59:23	9,524,071	4:42:07	1:04:41	868,039
Marathon year	2006.21	6.08	9,524,071	2009.19	2.49	868,039
Age	39.17	11.45	5,773,304	41.26	14.66	671,566
Male (1 = Male, 0 = Female)	0.67	0.47	8,714,067	0.61	0.49	806,518
Split 10km (minutes)	1:02:26	0:17:59	2,056,712	1:01:56	0:13:15	868,039
Split half marathon (minutes)	2:09:26	0:28:22	3,238,026	2:11:55	0:29:08	868,039
Split 30km (minutes)	3:12:36	0:44:55	1,496,139	3:12:35	0:43:47	868,039
Split 40km (minutes)	4:25:16	1:01:13	1,034,484	4:26:36	1:01:28	868,039

NOTE: The full splits sample includes marathons from the total marathon 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.

Table 2: Summary of Chetty et al. (2011) test for excess mass

	Full Sample				Sample with non-missing Age and Gender	Correcting for Boston Marathon Qualifiers	
Reference Point	Actual Finishers	Counterfactual Finishers	% Excess Finishers	T-Statistic	% Excess Finishers	% Excess Finishers	T-Statistic
3:00	86,770	69,661	24.6%	41.2	22.8%	22.8%	33.5
3:10	111,394	104,873	6.2%	14.7	7.2%	5.0%	6.6
3:20	159,648	158,006	1.0%	3.4	1.5%	2.2%	4.5
3:30	251,319	226,144	11.1%	46.7	10.6%	10.0%	22.2
4:00	408,322	360,959	13.1%	55.1	13.7%	13.8%	40.4
4:30	308,700	295,222	4.6%	19.8	4.5%	4.7%	16.4
5:00	212,986	202,039	5.4%	17.1	6.0%	5.9%	16.2
6:00	62,723	60,643	3.4%	6.2	3.5%	3.5%	6.1

NOTE: The correction for Boston Marathon omits the sub-sample 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.

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	
	# of	Mean Finishing	# of	% Excess		% Excess		% Excess	
Data Restriction	Marathons	Time	Finishers	Finishers	T-statistic	Finishers	T-statistic	Finishers	T-statistic
Year									
≤ 1990	48	234.87	328,467	17.4	11.2	13.7	10.3	9.9	4.2
1991 - 2000	234	250.58	841,280	21.4	13.4	13.1	17.1	6.4	8.0
2001 - 2010	4321	269.88	5,759,937	25.3	28.5	12.6	41.7	5.1	15.1
2011 - 2013	2228	267.43	2,594,387	27.0	27.1	14.2	34.3	5.5	10.0
# of Finishers									
> 10,000	201	272.25	4,395,209	26.1	34.8	12.8	41.0	4.5	11.1
5,000 - 10,000	219	257.45	1,524,712	27.4	22.8	12.3	22.8	4.6	6.1
1,000 - 5,000	1077	263.33	2,350,736	23.7	20.5	14.9	31.9	6.6	11.1
200 - 1000	2066	264.47	962,183	15.3	9.4	12.7	17.1	7.0	7.9
< 200	3268	275.65	291,231	14.9	4.9	8.9	6.1	8.4	4.6
Mean Marathon Finishing Time									
< 4:00	670	230.80	1,202,840	26.9	28.9	13.7	22.2	6.2	5.6
4:00 - 4:30	3594	256.66	5,229,502	24.6	32.4	13.7	52.0	5.2	14.0
> 4:30	2562	298.58	3,091,729	18.9	12.9	11.1	28.5	5.7	13.2
Region									
U.S.	5303	275.78	6,237,790	19.7	27.8	12.5	47.8	5.8	17.2
Europe	561	249.36	2,678,539	30.8	30.4	14.4	32.2	4.4	6.3
Canada	570	253.37	304,951	23.3	7.0	11.2	11.2	4.3	2.5
Other	401	256.03	302,791	21.9	8.5	14.2	13.9	5.8	4.2
Age									
≤ 29	--	272.13	1,262,936	20.7	13.4	13.6	19.2	4.5	6.0
30 - 39	--	264.72	1,864,056	21.8	17.2	13.7	22.9	5.8	10.2
40 - 49	--	266.04	1,657,318	26.3	18.6	13.8	23.5	6.5	8.9
≥ 50	--	266.87	1,223,081	24.6	10.6	12.6	17.9	6.4	9.0

NOTE: Analysis uses the total marathon sample ($n = 9,524,071$). The data restrictions divide marathon-years by year; the number of finishers; the average finishing time; and location. The excess mass measure is based on Chetty et al. (2011) test.