Marketing Faculty Publications   Marketing

2018

Goals as Reference Points in Marathon Running: A Novel Test of Reference Dependence

Alex Markle
University of Pennsylvania, amarkle1@wharton.upenn.edu

George Wu
University of Chicago, wu@chicagobooth.edu

Rebecca White
University of Chicago, rebecca.white@chicagobooth.edu

Aaron M. Sackett
sackett@stthomas.edu, sackett@stthomas.edu

Follow this and additional works at: https://ir.stthomas.edu/ocbmktgpub

Part of the Marketing Commons

Goals as Reference Points in Marathon Running: A Novel Test of Reference Dependence[*]

Alex Markle[†]   George Wu[‡]   Rebecca White[§]   Aaron Sackett[¶]

January 18, 2018

Abstract

In a large-scale field study of marathon runners, we test whether goals act as reference points in shaping the valuation of outcomes. Theories of reference-dependent preferences, such as Prospect Theory, imply that outcomes that are just below or just above a reference point are evaluated differently. Consistent with the Prospect Theory value function, we find that satisfaction as a function of relative performance (the difference between a runner's finishing time goal and her actual finishing time) exhibits loss aversion and diminishing sensitivity in both predictions of and actual experienced satisfaction. However, in contrast to Prospect Theory, we observe that loss aversion is partially driven by a discontinuity or jump at the reference point. In addition, we find that a runner's time goal as well as their previous marathon times simultaneously impact runner satisfaction, providing support for the impact of multiple reference points on satisfaction.

Keywords reference dependence, Prospect Theory, loss aversion, goals, satisfaction

JEL Classifications C93 D03

1   Introduction

A recent article in *The Chronicle of Higher Education* documented the enormous scholarly impact of Prospect Theory in economics, psychology, and beyond (Goldstein 2011). Perhaps the main reason Prospect Theory has been so influential is the theory's proposition that individuals evaluate outcomes in relative terms, as changes from a neutral reference point rather than final states as classical economic models assume (Kahneman and Tversky 1979; Tversky and Kahneman 1992).

Indeed, reference-dependent preferences account for a wide range of empirical phenomena, including the endowment effect (Kahneman et al. 1990), status quo bias (Samuelson and Zeckhauser 1988), disposition effect (Shefrin and Statman 1985), equity premium puzzle (Benartzi and Thaler 1995), and extreme risk aversion for prospects involving the possibility of a loss (Kahneman and Lovallo 1993; Rabin 2000).

Empirical investigations of Prospect Theory and reference dependence have largely taken the reference point to be the status quo.[1] In many domains, this choice is sensible because the status quo is the most reasonable or even the only reasonable reference point from which to evaluate a particular outcome. However, in other settings, points of comparison besides the status quo are plausible reference points, such as "aspirations, expectations, norms, and social comparisons" (Tversky and Kahneman 1991). While these alternative reference points have been the subject of some empirical study, such investigations face conceptual and methodological challenges. Conceptually, the field currently lacks the principles to guide a researcher challenged with identifying which of any number of plausible reference points is used to evaluate an outcome. For example, the reference point for selling a house might be the purchase price, an offer that was recently turned down, the sales price of a neighbor's property, a round number such as $500,000, or the average sales price of comparable houses. Empirically, even if the researcher knows which *type* of reference point drives evaluation, the *level* of that reference point may be difficult to observe and measure. Expectations, for example, are generally subjective and thus hidden from the researcher.

In this paper, we investigate reference dependence in marathon running.[2] Although this setting is distinct from those traditionally used to study Prospect Theory, we argue that it provides useful and unique insights into how non-status quo reference points operate. While people run marathons for a variety of reasons (e.g., to finish a long race or to support a charity), surpassing a predetermined time goal defines success for many marathoners. Indeed, marathon training guides often emphasize the importance of time goals as a motivational tool and provide heuristics and calculators for setting appropriate goals (e.g., Higdon 2011). In the present research, we test the specific proposition that

---

[1]Although we use reference-dependent preferences and Prospect Theory interchangeably, our investigation only considers the distortion of outcomes, not the nonlinear treatment of probabilities modeled by Prospect Theory (e.g., Wu and Gonzalez 1996; Prelec 1998).

[2]Some recent papers have used a combination of archival and survey data of marathoners to study the effect of round numbers on performance (Allen et al., 2017), optimism on performance (Krawcyzk and Wilamowski, 2017) and the relationship between performance and goal attainability (Burdina, Hiller, and Metz, 2017).

time goals serve as reference points in the evaluation of marathon performance.[3] In support of this proposition, we find that satisfaction as a function of performance *relative* to the time goal exhibits loss aversion and diminishing sensitivity, mirroring the classic S-shape of the Prospect Theory value function. However, we also document a jump in satisfaction at the reference point that departs from Prospect Theory's continuous value function. For marathon runners, aversion to losses thus appears to have two sources: a value function that has a steeper slope in losses than in gains, as in the classical Prospect Theory account, and a discrete jump at the reference point itself.

Our novel setting also yields a number of conceptual insights, filling gaps in the existing literature and generalizing the scope of Prospect Theory in several ways. First, we employ satisfaction as our dependent measure instead of estimating the value function from choice percentages or cash equivalents. While studies of overt behavior can speak to the role of reference dependence in the determination of *decision utility* (the weight given to outcomes in choice), they cannot provide insight into how, if at all, reference dependence shapes *experienced utility* (Kahneman 1999). In our study, we examine both pre-marathon predictions of satisfaction and post-marathon measures of satisfaction and find that loss aversion and diminishing sensitivity are present in both predicted and experienced satisfaction, contrary to the results of Kermer et al. (2006).

We also examine the role of multiple reference points. In our context, a time goal is one of several possible reference points runners may use to evaluate their performance. Experienced marathoners, for instance, may compare their current performance against their most recent or best marathon time. Indeed, we find that satisfaction among experienced marathoners is driven by comparing performance with both their time goal and previous marathon times. In sum, this study contributes to multiple strands of research on reference dependence.

Empirically, our results add to a growing list of field demonstrations of Prospect Theory (Camerer 2005, Camerer et al. 1997; Fryer et al. 2012; Genesove and Mayer 2001; Pope and Schweitzer 2011; Post et al. 2008; see Allen et al., 2017, for a more extensive list). This body of work suggests that reference dependence extends beyond small-stakes, stylized laboratory experiments, contrary to the criticism of List (2003).[4] Moreover, we provide an estimation of the

---

[3]The proposition that goals serve as reference points has broader implications. Heath, Larrick, and Wu (1999) proposed that goals also influence effort provision, an account supported by Sackett et al. (2015), Allen et al. (2017), and Corgnet et al. (2015).

[4]In the Electronic supplementary material, Section A.3.9, we also present evidence that loss aversion is present and of similar magnitude for both experienced and inexperienced marathoners, also contrary to List (2003).

Prospect Theory value function using data collected outside of the laboratory. Although many papers have estimated the Prospect Theory value function from laboratory data (e.g., Abdellaoui 2000; Gächter et al. 2007; Ho and Zhang 2008; Wu and Markle 2008), few have done so using field data (for exceptions, see Post et al. 2008; Tovar 2009).

Our paper proceeds as follows. In Section 2, we review the existing literature and position this paper's contributions relative to that literature. In Section 3, we describe the methods for our empirical investigation. In Section 4, we present the basic results of our study. In Section 5, we estimate the Prospect Theory value function to relate satisfaction to relative performance. In Section 6, we discuss some alternative explanations for our results and present evidence that casts doubt on these accounts. We conclude in Section 7, by summarizing our contributions and discussing some implications and future directions for this research.[5]

2 Background

In this section, we review literature on properties of the Prospect Theory value function, non-status quo reference points, the effect of multiple reference points on the valuation of outcomes, and differences between loss aversion in predicted and experienced utility. In doing so, we position the contributions of our investigation and suggest that our novel domain allows us to pose and answer questions that are difficult to address in more traditional decision-making settings.

2.1 Properties of the Prospect Theory Value Function

Theories of reference dependence imply that losses and gains are psychologically distinct. Behaviorally, the distinct treatment of gains and losses is captured by (i) the dislike for actuarially fair mixed gambles, such as a 50-50 chance to win or lose $1000; and (ii) the "reflection effect," the tendency toward risk aversion for gambles involving gains and toward risk seeking for gambles involving losses. These two phenomena led Kahneman and Tversky (1979) to posit a value function $v(x)$ that exhibits discontinuities in the first and second derivative at the reference point. The standard Prospect Theory value function is steeper for losses than gains and exhibits diminishing sensitivity, or concavity for gains and convexity for losses. Tversky and Kahneman (1992) fit a

---

[5]Additional details and analyses are found in the Electronic supplementary material.

4

value function to certainty equivalent data and provided empirical support for this S-shape.

Although the notion that "losses loom larger than gains" is intuitive, a precise definition of loss aversion is not straightforward. The literature has proposed a number of definitions of loss aversion, as well as corresponding indices (i.e., loss aversion coefficients) (Köbberling and Wakker 2005). In particular, loss aversion has been formalized either in terms of a difference in gain and loss *levels* $(-v(-x) > v(x), \forall x)$ or a difference in gain and loss *slopes* $(v'(-x) > v'(x), \forall x)$.[6] These definitions lend themselves to summary measures of loss aversion, $-v(-x)/v(x)$ and $v'(-x)/v'(x)$, and in turn reflect two behavioral propositions. In Prospect Theory, loss aversion in levels holds if an individual rejects all 50-50 symmetric gambles, whereas loss aversion in slopes holds if an individual prefers a 50-50 symmetric gamble to the same gamble with a mean-preserving spread.[7]

These two formulations are equivalent for a value function that is continuous at the reference point and exhibits symmetric curvature in gains and losses. However, a dislike of all symmetric gambles does not require loss aversion in slopes and could instead arise from other instantiations of reference dependence, such as a discontinuity or jump at the reference point as proposed by Diecidue and Van de Ven (2008). Similarly, a jump at the reference point does not imply an aversion to mean-preserving spreads captured by loss aversion in slopes. A discontinuous function thus distinguishes between loss aversion *at* the reference point and loss aversion *away from* the reference point.

We suggest that there is a parallel between the two behavioral propositions for mixed gambles and propositions about the motivating effect of goals. Heath, Larrick, and Wu (1999) suggested that a goal is motivating because it acts as a reference point. They asked participants for predictions of which of two tired individuals was more likely to do one and only one additional sit-up, varying goal levels as well as the number of sit-ups performed. When a goal is 30, it is intuitive that a person who has performed 29 sit-ups will be more motivated than someone who has already achieved 30 sit-ups. Such a difference in motivation reflects loss aversion *at* the reference point. However, Heath, Larrick, and Wu also found that a person who had performed 25 sit-ups was more motivated to do one more sit-up than someone who had performed 34 sit-ups, behavior consistent with loss aversion *away* from the reference point.

---

[6]Throughout, we assume that $v(0) = 0$.

[7]For rhetorical simplicity, we ignore the complication of a nonlinear probability weighting function (see Schmidt and Zank 2005).

These two examples provide support for two behavioral propositions about the motivating effect of goals on effort provision, at and away from the goal. In this paper, we investigate related propositions on the effect of goals on satisfaction. Loss aversion at a goal means that a one minute improvement in performance increases the satisfaction more for a runner who otherwise would have fallen one minute short of her goal than a runner who otherwise would have just achieved her goal, i.e., $v(0) - v(-1) > v(1) - v(0)$. Similarly, loss aversion away from a goal implies that the same improvement is more satisfying for a runner who is five minutes short of her goal than a runner who is four minutes ahead of her goal, i.e., $v(-4) - v(-5) > v(5) - v(4)$.

2.2 Non-Status Quo Reference Points

While studies of reference dependence have traditionally taken the reference point to be the status quo, an increasing number have examined alternative reference points. In studies of the disposition effect, for example, an asset's initial purchase price serves as the reference point from which later prices are evaluated. Consistent with that interpretation, investors are significantly more likely to sell winning stocks than losing stocks (Barberis and Xiong 2009; Genesove and Mayer 2001; Odean 1998; Shefrin and Statman 1985). Relatedly, Heath, Huddart, and Lang (1999) found that the yearly high price operates as the reference point in the evaluation of employee stock options, documenting a dramatic increase in the likelihood of exercising those options once the stock's price exceeded the yearly high.

Other studies have investigated contracts (Fehr, Hart, and Zehnder 2011; Hart and Moore 2008), fair wages and bonuses (Mas 2006; Ockenfels et al. 2014), and round numbers (Allen et al. 2017; Pope and Simonsohn 2011) as reference points. A great deal of research has investigated the role of expectations as reference points based on the theoretical models of Kőszegi and Rabin (2006, 2007) (cf. Abeler et al. 2011; Bartling, Brandes, and Schunk 2015; Ericson and Fuster 2011; Pope and Schweitzer 2011).

Goals, aspirations, and targets are natural candidates for reference points and have been incorporated into a number of theoretical choice models (Diecidue and Van De Ven 2008; Lopes 1987; Lopes and Oden 1999; March and Shapira 1992). In the literature on motivation, setting a goal is a "discrepancy creating process" that produces "discontent with one's present condition and the desire to attain an object or outcome" (Locke and Latham 2006). Implicit in this formulation is

6

the notion that goals operate as reference points, providing a point of comparison from which to evaluate present performance.

In a series of laboratory studies, Heath, Larrick, and Wu (1999) demonstrated that characteristics of the Prospect Theory value function could explain both effort and satisfaction in goal pursuit tasks. Camerer et al. (1997) found that New York City taxi drivers work longer hours on slower days (when the hourly wages are lower) and finish work earlier on busier days (when hourly wages are higher). They proposed that drivers establish daily income targets and compare their accumulated earnings against this standard. Earnings below a target is judged to be a loss, leading drivers to exert additional effort to meet their targets. Fehr and Goette (2007), Crawford and Meng (2011), Farber (2008), and Oettinger (1999) offered further support that income targets act as reference points, although Farber (2005) provided conflicting evidence.

The aforementioned studies highlight how non-status quo reference points can be used to provide unique insights into many important economic and non-economic phenomena. Nevertheless, there are conceptual and methodological challenges to studying non-status quo reference points. Heath, Larrick, and Wu (1999) proposed that "whenever a specific point of comparison is psychologically salient, it will serve as a reference point." However, the literature has provided few guiding principles to help researchers discriminate between any number of salient candidate reference points. Indeed, Barberis (2013) suggested that identifying the appropriate reference point is one of the main challenges to applying Prospect Theory in field settings: "It is often unclear how to define precisely what a gain or loss is, not least because Kahneman and Tversky offered relatively little guidance on how the reference point is determined" (p. 178).

The difficulties extend beyond conceptual determination to measurement. While historical stock prices are readily observable and round numbers are generally round for everyone, expectations and aspirations are both heterogeneous and difficult to observe.[8] Researchers have dealt with this complication by experimentally manipulating reference points as in Ericson and Fuster (2011) or Abeler et al. (2011), or inferring reference points from some observables as in Crawford and Meng (2011). In our study, we take a third approach, eliciting goals directly. Direct elicitation has the advantage that the reference point is known, but the disadvantage that participants are

---

[8]For instance, while Diecidue et al. (2015) found no support for an aspiration level at zero in a study on risky choice, both their data and those of Zeisberger et al. (2012) suggest heterogeneity in aspiration levels that may mask the ability to detest a jump at an aspiration level.

not randomly assigned to the reference point and therefore the inferred relationships may reflect systematic differences, for instance, between participants who set difficult goals and those who set easier ones. We address this complication directly in Section 6.

2.3   Multiple Reference Points

A student with a test score of "B+" could compare that result against her typical grade (say, a "B-"), her expected grade for that test (say, an "A-"), or the performance of her friend (say, a "B"). Which comparison will drive her evaluation of her performance on the test? In some cases, one of these reference points may be distinctly more psychologically salient than the others. But what happens if this is not the case? Much of the existing literature on multiple reference points has assumed that the reference point is some weighted average of a set of reference points (e.g., Briesch et al. 1997; Ordóñez 1998; Winer 1986) or that individuals direct their attention to only one reference point at a time (Lopes and Oden 1999; March and Shapira 1992). In contrast, Kahneman (1992) proposed that each of the possible reference points could *simultaneously* influence the evaluation of an outcome. In other words, an outcome could produce genuine mixed feelings or ambivalence, being viewed at once as a gain relative to one reference point and a loss relative to another (Larsen et al. 2004).

A handful of empirical studies have found support for Kahneman's simultaneous reference point account. Ordóñez et al. (2000) asked participants to rate their satisfaction with a hypothetical salary offer displayed alongside offers to other qualified job candidates. They documented a pattern of satisfaction consistent with the two reference points operating separately – lower satisfaction when the focal salary lay between two peer salaries than when the peers were offered the same salary as the subject. Sullivan and Kida (1995) found that managers simultaneously compared an investment's return to the return of previous investments and to a target return level. Finally, Wang and Johnson (2012) developed and provided empirical support for a tri-reference point extension to Prospect Theory that included both the status quo and a goal level as well as a minimum requirement (see also Weingarten et al. 2016).

Like the examples above, a marathon runner might compare her performance against her time goal, her most recent marathon time, her personal best marathon time, or a number of other possible points of comparison. While we hypothesize that comparisons between one's actual performance

and time goal will be the primary driver of a runner's satisfaction, we also test whether comparisons with previous performance also influence satisfaction.

2.4 Loss Aversion in Predicted and Experienced Utility

"Utility" as most commonly used in economics and decision making research today refers to the weight given to an outcome as inferred from observed choices. This usage is distinct from Bentham (1789), who took utility to be the actual affective experience (pain or pleasure) aroused by an outcome. Kahneman et al. (1997) termed the earlier notion of utility *experienced utility* and the later and more common conception *decision utility.* In our context, there are no decisions so we refer to predictions of utility as *predicted utility.* There is significant evidence that experienced and predicted utility do not necessarily coincide. People often fail to correctly anticipate and predict their affective reactions to events (Gilbert et al. 1998; Wilson and Gilbert 2003). Furthermore, the prevailing finding is of an *impact bias* or systematic overestimation of both the intensity and duration of emotional reactions (Wilson, Myers, and Gilbert 2003). Of particular relevance for this study, Kermer et al. (2006) argued that individuals overestimate the hedonic impact of losses, suggesting that loss aversion is a type of affective forecasting error (see, however, Boyce et al. 2013).

The Prospect Theory value function is explicitly a representation of the predicted utility of outcomes, rather than the actual hedonic experience of those outcomes (Kahneman 1999). Whether the characteristics of the value function such as loss aversion also apply to experienced utility is largely an open question. Studies of choice often implicitly assume an equivalence between experienced and decision utility, but choice data can only reveal the decision utility of an outcome. Our study elicits both predicted satisfaction before the marathon and actual satisfaction after the marathon and thus allows for a test of reference dependence in both predicted and experienced utility.

3 Methods

3.1 Participants

We recruited 2,652 participants registered for one of 15 targeted marathons conducted from 2007 to 2009.[9] Each marathon was among the 20 largest U.S. marathons, with the number of finishers ranging from 6,875 for the 2008 Grandma's Marathon to 38,557 for the 2007 New York Marathon.[10] Participants were compensated by being entered into a random drawing for prizes. The Electronic supplementary material contains a more complete description of the study, including the wording and order of all survey questions (Section A.1) and summary statistics of the demographic and goal measures (Section A.2).

We dropped participants from our sample who did not complete the entire study (23.5%), who did not start (2.3%) or finish (4.6%) the marathon they entered, whose data could not be matched to official marathon results (2.5%), or who had participated in a previous targeted marathon (0.6%). These exclusions left us with 1,801 runners. Our sample was reasonably representative of the overall population of marathon finishers used in our study. We use the 283,651 marathon finishers in our 15 sampled marathons as the "population" and the 1,801 runners described above as our "sample."[11] Our sample ($M = 274.99$) had slightly faster finishing times than the population of marathoners ($M = 279.41$). The mean age of our runner sample, 37.33, was almost identical to the mean age of runners in the marathon population, 37.27. However, our sample had more women ($M = 57.3\%$) than the population ($M = 40.6\%$), a pattern consistent with the general finding that women are more likely to participate in surveys than men (Curtin, Presser, and Singer 2000).

3.2 Conditions

Participants were randomly assigned to one of six conditions created by crossing three pre-marathon conditions with two post-marathon conditions. Participants in our study were blind to condition

---

[9]The marathons surveyed were Boston (2008), Chicago (2007-2009), Grandma's (2008), Los Angeles (2008), Marine Corps (2007-2009), New York City (2007), Portland (2007), Rock 'n' Roll San Diego (2008), and Twin Cities (2007-2009).

[10]`http://www.marathonguide.com/Features/Articles/2007RecapOverview.cfm`. Referenced on January 18, 2018.

[11]We created weighted averages by weighting the relevant statistics by the proportion of our sample in each marathon. For example, 18.4% of our participants ran the 2009 Marine Corps Marathon. To compute the weighted average finishing time, we multiplied the finishing time for all runners in the 2009 Marine Corps Marathon (281.23 minutes) by 18.4%, repeated this process for the other 14 marathons, and summed the 15 products.

and were unaware of our specific research hypotheses other than that we were examining the "relationship between marathon performance and satisfaction." Although we describe the pre- and post-marathon conditions in detail, we collapse the conditions in our estimation of the value function. Results by pre- and post-marathon condition are found in the Electronic supplementary material, Sections A3.7 and A3.8. All surveys were administered online.

3.2.1 Pre-Marathon Conditions

The pre-marathon conditions consisted of a "goal-not-asked" ($n = 746$), an early "goal-asked" ($n = 331$) and a late "goal-asked" ($n = 724$) condition. The separate timing conditions were employed for reasons not related to this paper, and we simplified the design for the 2009 marathons, dropping the early goal-asked condition.

Participants in the goal-not-asked condition were surveyed approximately two weeks prior to the marathon and asked to provide some basic demographic information, including their age, gender, and information on their training and running background. Runners indicated whether they had previously completed a marathon, and provided their fastest and most recent marathon times, and their fastest half-marathon and 10-kilometer times. Participants in this condition were not asked about goals or any other objectives they might have had for their upcoming marathon.

Participants in the early goal-asked condition completed two surveys prior to the marathon, administered approximately 6 and 2 weeks before the marathon, while participants in the late goal-asked condition completed only one survey, administered approximately 2 weeks prior to the marathon. We asked participants in both of the goal-asked conditions to report the same demographic and experience information as the goal-not-asked participants. In addition, participants in the goal-asked conditions were asked about time goals and other objectives. Specifically, participants were asked if they had a "specific time goal for the marathon," and if so, to indicate it. They also estimated the likelihood of reaching the goal and were asked to predict their satisfaction on a 1-to-7 scale for finishing 1, 10, and 20 minutes ahead of or behind their time goal and to indicate on a 1-to-7 scale the importance of meeting their time goal. Participants in the early condition answered these questions twice, while participants in the late pre-condition answered these questions once. We included both "goal-asked" and "goal-not-asked" pre-marathon conditions to test whether providing a goal prior to the marathon increased the commitment to that goal, possibly

leading to increased loss aversion, as well as to test whether providing a goal prior to the marathon affected performance (for more details, see Sackett et al. 2015).

3.2.2 Post-Marathon Conditions

The two post-marathon conditions differed only in timing. Participants in the early ($n = 1,183$) condition were contacted 1 day after the completion of their marathon, while participants in the late ($n = 618$) condition were contacted 4 weeks after the marathon finished. In both cases, participants had 2 weeks to complete the survey. We omitted the late post-condition for the 2009 marathons. We included the two conditions to test whether loss aversion was attenuated over time, as might be predicted by some accounts of affective forecasting (Gilbert et al. 1998).

Participants were asked if they started the marathon, and if so, whether they completed the race. Runners who completed the marathon then rated their satisfaction with their performance as well as effort on a 1-to-7 Likert scale. Participants next rated a number of goals, including meeting their time goal, on importance on the same scale. Participants then indicated their time goal if they had one. Note that participants in the goal-asked pre-marathon conditions had previously provided a goal, while participants in the goal-not-asked condition had not. Critically, to minimize the possibility that any observed reference-dependent satisfaction was cued by the experimenter, participants provided their time goal and marathon finishing time *after* they had completed the satisfaction question as well as a number of other questions about the importance of several possible objectives.[12]

Participants also reported their official time, as well as their bib number. We used these to match their reported results with their official results.

3.3 Official Performance Data

We matched participant responses with official marathon results obtained from marathon websites. Large marathons such as the ones in our study typically provide two times, a "chip time" (also known as "net time") and a "clock time" (also known as "gun time"). The chip time reflects the elapsed time between when a runner crosses the start line and the finish line, while the clock time is

---

[12]Of course, there is no obvious way to make satisfaction ratings incentive-compatible. Nevertheless, most studies of well-being use self-reports of happiness and satisfaction (cf. Schwarz and Strack 1999).

the difference between when a runner finishes the race and when that race officially starts. For large races like the ones in our sample, the difference between the clock and chip time can be sizable. Our runners took an average of 7.86 minutes to reach the starting line. We use the official chip time throughout as a measure of performance, as this is the number that most runners take to be the most relevant measure of performance (Austen 2001).[13]

4 Basic Results

In this section, we describe some basic results. In Section 5, we organize these results by estimating a Prospect Theory value function to relate satisfaction with relative performance.

4.1 Time Goals

For all of the analyses that follow, we use goals that are elicited post-marathon to maximize our sample size. (In the Electronic supplementary material, Section A.3.6, we report conceptually similar results using pre-marathon goals for the goal-asked participants.) Overall, 86.1% or 1,550 of our 1,801 participants had time goals. Experienced marathoners ($M = 88.2\%$) were more likely to have goals than were rookie marathoners ($M = 78.8\%$) ($\chi^2(1) = 23.65$, $p < .001$). 60.7% of experienced marathoners set goals faster than their best marathon time, with 82.6% setting goals faster than their last marathon time.

We asked participants to indicate the importance of attaining their time goal on a 1-to-7 scale. Runners who provided a goal rated that goal as being important on average ($M = 5.80$), with 86.4% of runners rating the goal as 5 or higher in importance. Participants were also optimistic about the chances of reaching their goal, indicating a 72.0% chance of achieving their goal on average.

Participants in the goal-asked conditions provided goals prior to and after the marathon. These marathoners exhibited a small but statistically significant tendency to report less ambitious goals after the marathon ($M = 246.50$) than before it ($M = 245.23$) (paired t-test, $t(835) = 3.30$, $p < .001$). We provide more details about how goals change in the Electronic supplementary material, Section A.3.6.

---

[13]The chip time is used as a qualifying time for "elite" races such as the Boston Marathon. However, clock time is the time generally used for determining prize money and other awards.

In the remainder of the paper, we restrict our analysis to participants who provided a time goal, a post-race satisfaction measure, a measure of goal importance, and the number of marathons they had previously run. These exclusions leave us with 1,534 observations.

4.2   Performance

The average finishing time for our runners was 4:29:70 (male: 4:12:40; female: 4:41:53), with finishing times ranging from around 2:15 to over 8 hours.[14] 24.7% of our runners bettered their time goal, with the success rate varying considerably across marathons, ranging from 3.3% for the 2007 Chicago Marathon to 45.8% for the 2009 Twin Cities Marathon. 35.4% of runners who rated achieving their time goal as 7 out of 7 in importance met their goal, compared to 18.7% of all remaining runners ($\chi^2(1) = 52.20$, $p < 0.001$). In addition, 23.4% of experienced (i.e., non-rookie) marathoners beat their time goal, compared to 28.5% of first-time marathoners ($\chi^2(1) = 3.90$, $p = .098$).
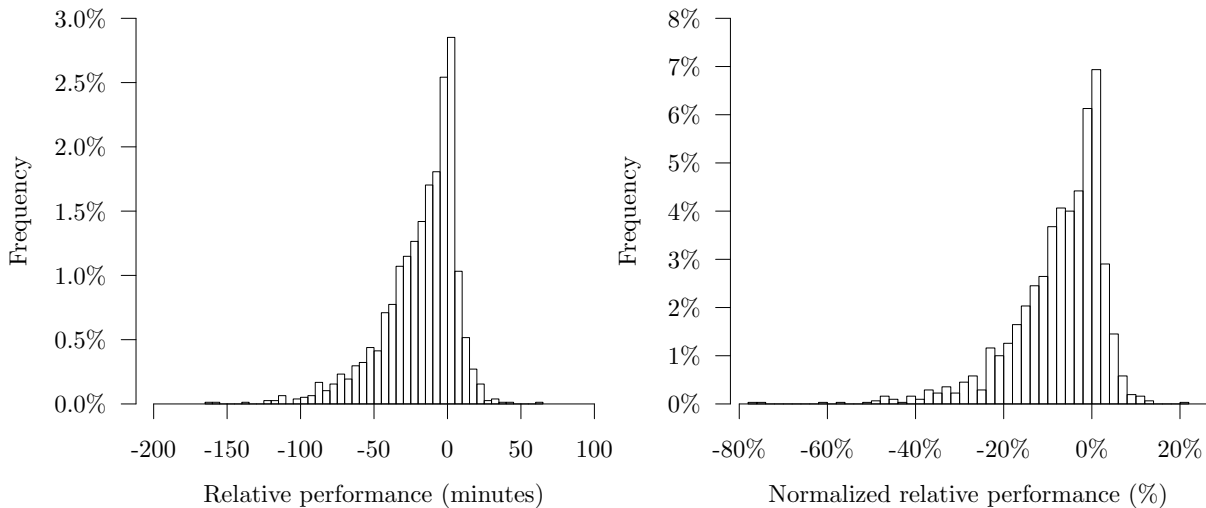
Figure 1: Histograms of finishing time relative to time goal. The left panel plots relative performance in minutes with 5 minute bins. The right panel plots normalized relative performance in % with 2% bins.

We define "relative performance" to be the difference between finishing time and time goal. Runners fell short of their goal by 18.27 minutes on average (Median: 11.84). To deal with the large variation of finishing times, we also created a "normalized relative performance" measure by

---

[14]Precise finishing times are not provided to protect the anonymity of our participants.

dividing relative performance by time goal. For example, a runner who ran 4:12 and had a goal of running 4:00 was 5% (12/240) short of her goal. The mean normalized relative performance was 6.1% short of the goal (Median: 4.7%). Figure 1 shows histograms of both measures of relative performance.

4.3   Predicted Satisfaction

We asked participants in the goal-asked conditions to predict their satisfaction from finishing 1, 10, and 20 minutes ahead or behind their goal. The 20 minute questions were added for the 2008 and 2009 surveys. Thus, this analysis only uses 2008 and 2009 participants in the goal-asked conditions ($n = 613$). The average of these predictions is shown in Figure 2, where positive values along the x-axis indicate performance that betters the goal.

Note that this figure traces the characteristic shape of the Prospect Theory value function. Because the satisfaction scale is ordinal, we analyze these results with ordered logistic regression, predicting expected satisfaction $S(t)$ based on performance relative to the goal, t. Our model includes dummy variables for finishing t minutes ahead of the goal, I_{+t}, $t = 1, 10, 20$, and t minutes behind the goal, I_{-t}, $t = 1, 10, 20$, as well as fixed effects for each of the participants:

$$S(t) = n_{-20}I_{-20} + n_{-10}I_{-10} + n_{-1}I_{-1} + p_{+1}I_{+1} + p_{+10}I_{+10} + p_{+20}I_{+20}.$$
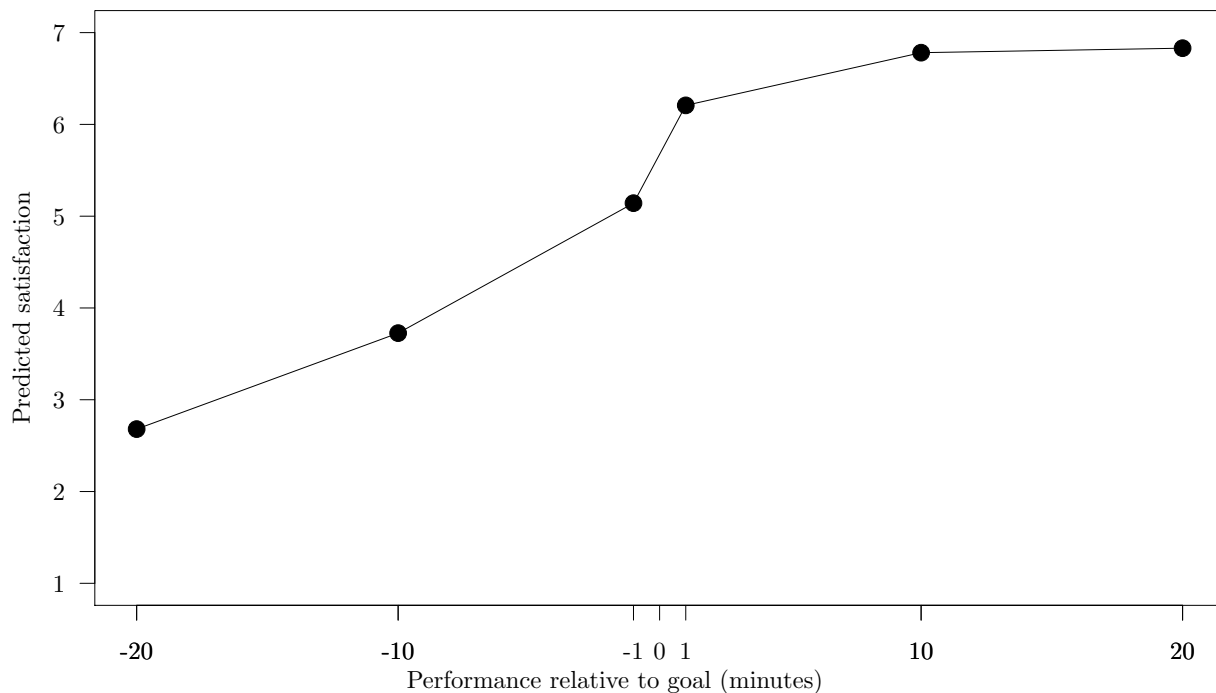
Figure 2: Predicted satisfaction as a function of relative performance.

This model reveals statistically significant levels of loss aversion and diminishing sensitivity. Although there are multiple ways to define and measure loss aversion (see Abdellaoui, Bleichrodt, and Paraschiv 2007), in our context, the most natural definition compares the effect on satisfaction of improving from 20 minutes short of the goal to 1 minute short of the goal, $n_{-1} - n_{-20}$, to the effect on satisfaction of exceeding the goal by 20 minutes rather than 1 minute, $p_{+20} - p_{+1}$. A Wald test shows that a decrease in losses increases satisfaction more than an increase in gains, $\chi^2(1) = 11.50$, $p < .001$. We test for diminishing sensitivity in losses by testing whether $.9(n_{-10} - n_{-20}) = n_{-1} - n_{-10}$ ($\chi^2(1) = 13.68$, $p < .001$) and in gains by testing whether $.9(p_{+20} - p_{+10}) = p_{+10} - p_{+1}$ ($\chi^2(1) = 26.02$, $p < .001$).[15]

4.4 Actual Satisfaction

Mean satisfaction with performance was 4.84 out of 7. Satisfaction was negatively correlated with finishing time, with faster finishers more satisfied than slower finishers (Spearman $\rho = -0.24$,

---

[15]Although a comparison of $n_{-10} - n_{-20}$ to $p_{+20} - p_{+10}$ reveals significant loss aversion ($\chi^2(1) = 14.83$, $p < .001$), a comparison of $n_{-1} - n_{-10}$ to $p_{+10} - p_{+1}$ does not ($\chi^2(1) = 0.57$, $p = .45$). In the Electronic supplementary material, Section A.3.10, we present two alternative analyses: a parametric analysis using paired t-tests and an ordered logit analysis using piecewise polynomials, as in Section 5.2.

$p < .001$). However, the correlation was much stronger if we considered performance relative to a runner's time goal, either relative performance in minutes (Spearman $\rho = -0.59$, $p < .001$) or normalized relative performance in percentage terms (Spearman $\rho = -0.60$, $p < .001$).

5 Estimation of the Value Function

In this section, we test whether the relationship between satisfaction and performance exhibits reference dependence. There are several challenges to uncovering such a relationship (Frey and Stutzer 2002). First, each participant provided only one satisfaction rating, and thus we must estimate an aggregate value function. Second, our satisfaction measure was elicited on a 1-to-7 Likert scale. Although such a measure is often interpreted as being a cardinal measure, it should be thought of as an ordinal scale (cf. Long and Freese 2014). Third, the discreteness and boundedness of this scale violates the standard regression assumption of normally distributed errors, making OLS or even nonlinear regression inappropriate. Finally, there is the challenge of measuring loss aversion. Köbberling and Wakker (2005) discuss the issues in defining loss aversion, in particular in cases in which the value functions for gains and losses do not have the same shape.

To deal with the first challenge, we assume that satisfaction is a function of normalized relative performance, the difference between the time goal and the finishing time divided by the time goal. This normalization makes performance across participants more readily comparable and appropriate for aggregation. We deal with the second and third challenge by employing ordered logit regression, which does not assume equal spacing between the categories of the response variable. Although an ordered logit analysis is statistically appropriate for our setting, the interpretation of coefficients is more complicated than in standard value function estimation exercises. Ordered logit regression models the relationship of some predictor variables (in our case, relative performance, as well as some control variables) and some latent variable (in our case, satisfaction), and then relates values of that latent variable to some ordinal scale (in our case, a 1-to-7 Likert rating). The regression coefficient of a predictor variable therefore does not provide an estimate of the direct effect on our measure of satisfaction, but instead indicates the effect of that predictor variable on the log odds of one satisfaction level relative to all inferior satisfaction levels. Accordingly, we perform inference and present the fitted model parameters on the scale of the linear predictors, but plot the

relationship both on the scale of the linear predictors as well as in the form of expected satisfaction on the original ordinal scale. Even though the latter plot violates the ordinal properties of the satisfaction scale, we provide this plot for heuristic purposes and because the scale is more readily interpretable (Harrell 2010).[16] Finally, as discussed in Section 2.1, we examine loss aversion by considering two measures: (i) a jump in satisfaction from reaching a goal; and (ii) the difference in slopes in the gain and loss domains away from 0.

We adopt multiple approaches to test for the main properties of reference dependence, loss aversion and diminishing sensitivity. We first fit a smoothing spline to the data to provide a general sense of the shape of the underlying relationship. Although a smoothing spline is useful for exploratory data analysis, it is not amenable to statistical inference. We therefore also fit the data with piecewise polynomials. We use these piecewise functions to provide a statistical test of whether satisfaction with performance exhibits reference dependence, looking for a jump in the value function, a difference in the gain and loss slopes away from 0, and diminishing sensitivity.[17] The two approaches, the smoothing spline and the piecewise polynomials, offer a consistent picture of reference-dependent preferences.

5.1 Smoothing Spline

We first provide a general sense of the relationship between relative performance and satisfaction by fitting a smoothing spline to the raw data (see Figure 3). While smoothing splines are not easily amenable to inference (because n data points are modeled by $n-1$ separate polynomials), they can provide insight into the shape of the underlying function without making any strong *a priori* assumptions about that shape (Green and Silverman 1994). Our smoothing spline is constructed from separate cubic polynomials between each data point.[18]

Smoothing splines impose continuity constraints on the first and second derivatives and thus

---

[16]For any combination of predictors, the ordered logit model produces a probability distribution for each of the 7 satisfaction levels. We can thus plot expected satisfaction for any level of relative performance. When we employ control variables such as dummy variables for a specific marathon, the control variables are set at their mean levels.

[17]The power function is the most commonly-used parametric form for estimating the Prospect Theory value function (e.g., Abdellaoui et al. 2007; Wu and Gonzalez 1996). It has a single parameter governing curvature and a single parameter indexing loss aversion when the power parameter for gains and losses coincides. However, parametric forms such as the power function cannot be readily incorporated within ordered logit models.

[18]The smoothing spline has a single parameter, which determines the tradeoff between the curve's smoothness and its closeness to the data. This parameter can be chosen subjectively or determined analytically by leave-one-out cross-validation, the approach taken here. Additional discussion of smoothing splines is found in Green and Silverman (1994) and Wang (2011).

are unable to capture the discontinuities at the reference point that are characteristic of reference-dependent preferences. Accordingly, the right panel of Figure 3 consists of two smoothing splines, fit separately for observations below and above the reference point. Both panels reflect Prospect Theory's characteristic S-shape, with the right panel also capturing both aspects of loss aversion discussed in Section 2.1, a discontinuity or "jump" at the reference point as well as a function that is steeper for losses than gains.
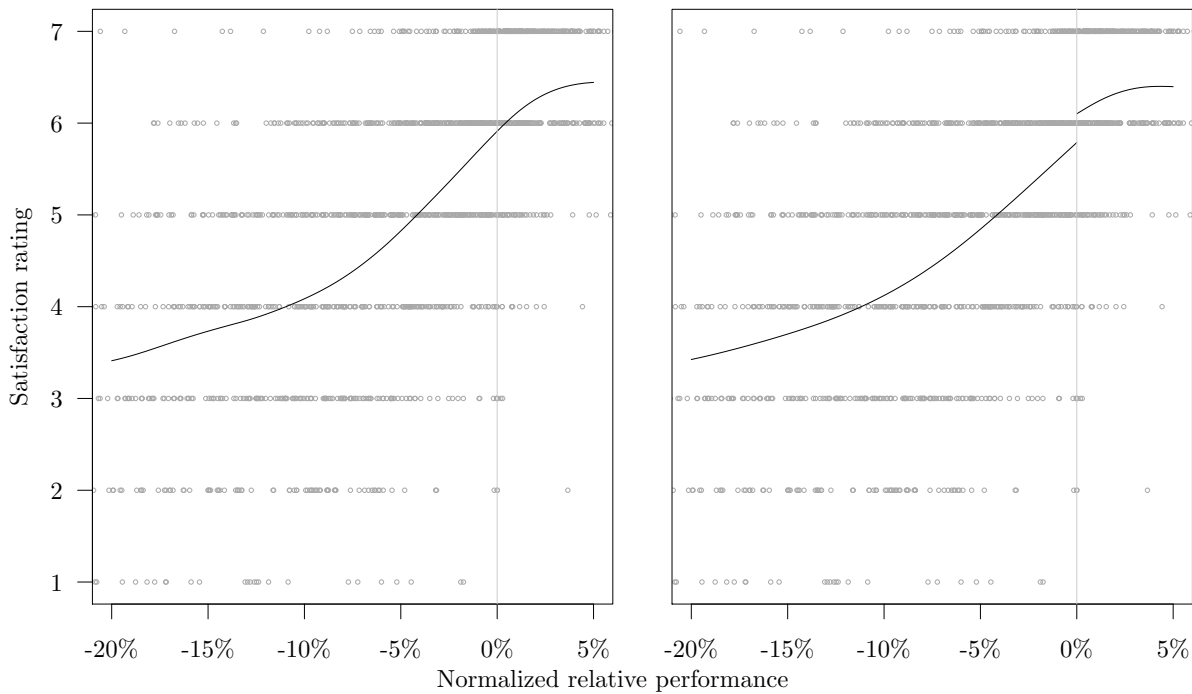
Figure 3: Raw data and a smoothing spline fit to the full domain of performance (left panel) and raw data and two smoothing splines, one fit below and the other above zero (right panel).

5.2 Piecewise Polynomial Regression Models

We next fit piecewise polynomial models using ordered logit regression to relate satisfaction with normalized relative performance. These models consist of two functions fit separately for negative and positive relative performance and allowing for a discontinuity at the reference point:

$$S(x) = \begin{cases} \sum_{i=1}^{a} n_i x^i, & x < 0 \\ \sum_{i=1}^{a} p_i x^i \ + J, & x \geq 0 \end{cases}.$$

19

The parameter a captures the order of the polynomials. We fit linear ($a = 1$), quadratic ($a = 2$), and cubic ($a = 3$) polynomials, including controls for each of our 15 marathons, as well as the gender of the participant. The quadratic model is favored by both the Bayesian Information Criterion (BIC) (linear: 4903; quadratic: 4864; cubic: 4868) and bootstrap validation.[19] For these reasons as well as for parsimony, we use the quadratic model for our main analyses. The Electronic supplementary material, Section A.3.1, shows that our main results described below hold for polynomials besides the quadratic, while Section A.3.2 provides validation for splitting the piecewise polynomials at 0, our presumed reference point.

Parameter estimates for the piecewise quadratic polynomial are presented in Table 1. Ordered logit parameters are on a log odds scale, and thus are commonly interpreted in terms of odds ratios, which for a parameter estimate $\hat{\beta}_k$ equals $\exp(\hat{\beta}_k)$. A one unit increase in k changes the odds of a higher response (for instance, a 7 versus a 1 through 6) by a factor of $\exp(\hat{\beta}_k)$, holding all other variables constant. For example, a 1% increase in performance from 0% to 1% increases the odds of reporting higher satisfaction by a factor of $\exp(\hat{p}_1(.01) + \hat{p}_2(.01)^2) = 1.26$.[20] The left panel of Figure 4 plots the linear predictors from the quadratic model on a log odds scale, while the right panel plots expected satisfaction on the original ordinal scale.

---

[19] Our bootstrap validation involves drawing a bootstrap sample with replacement from the original sample, then fitting the model to be validated to the bootstrap sample. That fitted model was then applied to the original sample, and Somers' d, a measure of the association between predicted probabilities and observed responses, was compared against that obtained by fitting the model directly to the original sample. This provided an estimate of the bias due to overfitting, also called the model's "optimism." This process was repeated 1000 times, and the average optimism was subtracted from the index of accuracy from the original sample, producing an overfitting-corrected estimate (Harrell 2010).

[20] In the proportional odds model, the inference is independent of the reference category, i.e., the effect of an increase in variable k on the log odds of $\Pr(Y \geq j)$ is the same for any level of the response variable, j.

	Predictor	$\hat{\beta}$
Linear terms	losses $(\hat{n}_1)$	18.06***
	gains $(\hat{p}_1)$	24.57***
Quadratic terms	losses $(\hat{n}_2)$	19.53***
	gains $(\hat{p}_2)$	-181.62***
	jump $(\hat{J})$	0.97***
Observations		1534
degrees of freedom		20
-2 log likelihood		4673.04

Table 1: Ordered logit regression on piecewise polynomials relating satisfaction to normalized relative performance. The statistical significance of coefficients is indicated by: *** ($p < .01$), ** ($p < .05$), and * ($p < .10$).
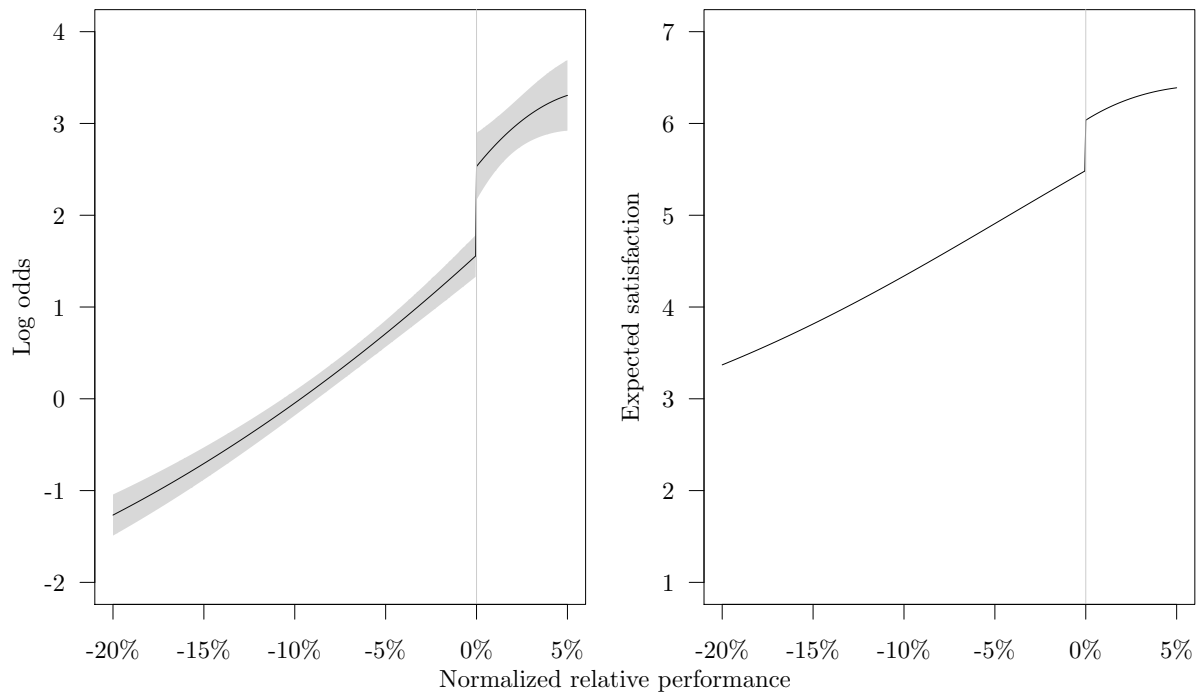

Figure 4: Ordered logit regression on piecewise quadratic polynomial relating satisfaction to normalized relative performance. The left panel plots the log odds of meeting or exceeding the midpoint of our satisfaction scale (4), along with a 95% confidence band, while the right panel presents expected satisfaction on the original ordinal (1-7) scale.

5.2.1   Test of reference dependence

Loss aversion

In Section 2.1, we distinguished between loss aversion *at* the reference point (a jump), and loss aversion *away from* the reference point (a difference in slopes). We begin by considering loss aversion at the reference point. The jump at the reference point is significant and positive, $\chi^2(1) = 24.89$, $p < .001$. Just meeting one's goal more than doubles the odds of reporting higher satisfaction, $\exp(0.97) = 2.63$.

We next consider loss aversion away from the reference point. Loss aversion is commonly measured as the ratio of the loss and gain slopes or loss and gain levels. When the value function is determined from certainty equivalents for gambles, these ratios have a cardinal interpretation, inspiring the common observation that "losses are about twice as painful as gains are pleasurable" (Hastie and Dawes 2001, p. 216). Because the ordered logit model maps the observed categories onto a continuous latent variable, effects of that variable can only be interpreted ordinally. Ordered logit coefficients reflect the log odds of reporting satisfaction greater than or equal to any given level, therefore even the ratio $\hat{n}_1/\hat{p}_1$ in the piecewise linear model ($a = 1$) has no straightforward interpretation and thus is *not* a natural measure of loss aversion. As a result, we do not provide measures of loss aversion but instead test whether there is a statistically significant difference between the loss and gain slopes. A second complication with polynomial functions is that the loss and gain slopes vary with the distance x from the reference point. We deal with this complication by comparing the gain and loss slopes for different distances x from the reference point, $S'(x)$ and $S'(-x)$, ranging from 0.5% to 5%. We choose this cutoff because only 3.2% of our data for gains exceeds 5%. The results of this analysis are presented in Table 2 for the best fitting quadratic model.

As we move away from the reference point, the slopes in both the loss and gain domains get smaller (see the section on diminishing sensitivity below), but they do so at different rates. The slope for losses exceeds the slope for gains for $x \geq 2.1\%$ and becomes significantly steeper than the slope for gains (at the 0.05 level) for $x \geq 4.6\%$. There are conceptual, statistical, and measurement reasons that we find modest amounts of loss aversion on this measure, contrary to Booij and van de Kuilen (2009) and others. First, while a conceptual distinction can be made between the jump

at the reference point and the difference in slopes in the gain and loss domains, they are difficult to disentangle statistically. This is especially difficult with higher order polynomials, for which a steeply curved loss domain can subsume the discontinuity at the reference point. Indeed, we observe a more steeply sloped function in the loss domain for the higher ordered models, coupled with smaller estimates for the size of the jump at the reference point. For example, for a cubic model, the loss slope is significantly greater than the gain slope for $x \geq 3.7\%$ and the region in which loss aversion in slopes is significant grows as the order of the polynomial in losses increases (see Electronic supplementary material, Section A.3.1). Second, the confidence intervals for small losses or small gains are wider than for more intermediate losses or gains (see Figure 4), reflecting the general tendency for confidence intervals to be narrowest at intermediate levels (e.g., Draper and Smith 1966). Finally, our relative performance measure reflects measurement error because of a linguistic ambiguity in how goals are communicated. We asked our participants to provide a time goal. A runner with a goal to finish the marathon in g minutes can express the goal as g or $g - 1$ (1 second faster than g). For example, Allen et al. (2017) documented the tendency of marathon finishing times to bunch just below round numbers such as 4 hours. In our sample, 179 of our participants provided a goal of 4:00:00, 58 participants a goal of 3:59:59, and 13 participants a goal of 3:59:00. We suspect that most of the latter participants were communicating that their goal was to run in "3:59 something." Indeed, only one runner provided a goal of 3:58:00.

Finally, we compare the gain and loss *levels*, $S(0) - S(-x)$ and $S(x) - S(0)$, at different distances x from the reference point. This comparison incorporates the jump at the reference point and the difference in slopes away from the reference point. Loss levels are significantly greater than gain levels for all values of x from 0% to 5%.

Diminishing sensitivity

Diminishing sensitivity away from the reference point is equivalent to $n_2 > 0$ for losses and $p_2 < 0$ for gains. We find strong support for diminishing sensitivity in both domains, losses, $\chi^2(1) = 45.57$, $p < .001$, and gains, $\chi^2(1) = 9.38$, $p = .002$.

	Slopes			Levels		
x	loss	gain	p-value	loss	gain	p-value
0.0%	18.06	24.57	0.414	0.97	0.00	0.000
0.5%	17.87	22.76	0.512	1.06	0.12	0.000
1.0%	17.67	20.94	0.639	1.15	0.23	0.000
1.5%	17.48	19.12	0.800	1.23	0.33	0.001
2.0%	17.28	17.31	0.997	1.32	0.42	0.002
2.5%	17.09	15.49	0.776	1.41	0.50	0.005
3.0%	16.89	13.68	0.536	1.49	0.57	0.008
3.5%	16.70	11.86	0.318	1.57	0.64	0.011
4.0%	16.50	10.04	0.154	1.66	0.69	0.013
4.5%	16.31	8.23	0.059	1.74	0.74	0.014
5.0%	16.11	6.41	0.018	1.82	0.77	0.013

Table 2: Tests of loss aversion in the relationship between satisfaction and normalized relative performance. The table presents loss and gain slopes ($S'(-x)$ and $S'(x)$) and loss and gain levels ($S(0) - S(-x)$ and $S(x) - S(0)$) at different values of x for the quadratic model. The p-values are obtained by Wald tests that examine the null hypothesis of equality of the loss and gain slopes and the loss and gain levels, respectively.

5.2.2 Robustness Analysis

We report robustness analyses in the Electronic supplementary material, Section A.3. We show there that the main results on loss aversion of levels and diminishing sensitivity are largely robust to the choice of order of loss and gain polynomials (A.3.1), the use of relative performance in minutes instead of normalized relative performance as a measure of performance (A.3.3), and the exclusion of "influential" outliers (A.3.4).

5.2.3 Alternative reference points

We have assumed so far that goals serve as reference points. However, it is possible that other comparisons that are correlated with goals are really driving the relationship between performance and satisfaction. Therefore, we contrast our goals-as-reference-points model with two alternative models: one in which performance is defined relative to a runner's best marathon time and one in which performance is defined relative to a runner's most recent marathon time. Not surprisingly, these two measures are highly correlated with performance relative to time goal (Best Marathon: $r = 0.42$; Last Marathon: $r = 0.39$). Because the latter two reference points are only applicable to runners with previous marathon experience, we restricted this analysis to experienced runners who reported both a best and most recent marathon time ($n = 1,015$). The regression coefficients for these analyses are presented in Table 3.

Although all three models have similar shapes, the model with time goal as a reference point is clearly the best fitting model as measured by log likelihood. Correspondingly, the coefficients for the best and last marathon time models are considerably smaller than for the model that uses time goal as a reference point. For example, bettering a goal by 1% instead of 0% increases the odds of reporting higher satisfaction by a factor of 1.37. The same factors for best and last marathon are 1.04 and 1.05, respectively. Critically, bettering one's best or last marathon produces a significant jump in satisfaction (Best: $\chi^2(1) = 14.59$, $p < .001$; Last: $\chi^2(1) = 13.10$, $p < .001$). However, neither model exhibits a significant difference in slopes between the gain and loss domains, either at the reference point or at $x = 5\%$, or diminishing sensitivity in either gains or losses.

		Reference Point		
	Predictor	Time Goal	Best Time	Last Time
Linear	losses $(\hat{n}_1)$	19.61***	2.65***	2.81*
	gains $(\hat{p}_1)$	35.11**	3.84	4.99**
Quadratic	losses $(\hat{n}_2)$	21.18***	1.67	4.24
	gains $(\hat{p}_2)$	−329.79	−9.16	−6.83
	Jump $(\hat{J})$	0.77***	0.82***	0.77***
Observations		1015	1015	1015
degrees of freedom		20	20	20
-2 log likelihood		3102.18	3423.30	3406.08

Table 3: Alternative reference point analysis. Ordered logit regression on quadratic polynomial relating satisfaction to normalized relative performance, defined relative to three alternative reference points: a runner's time goal, their best previous marathon time, and their most recent previous marathon time. The statistical significance of coefficients is indicated by: *** (**$p < .01$**), ** (**$p < .05$**), and * (**$p < .10$**).

5.3 Multiple Reference Points

The previous analysis suggests that comparisons between performance and time goal drive satisfaction to a greater degree than do comparisons between performance and past marathon times. Nevertheless, this analysis does not preclude the possibility that the latter comparisons additionally influence their satisfaction as in Kahneman's (1992) multiple reference point account. Therefore we test whether incorporating multiple reference points in the model contributes to its explanatory power. Let x_g and x_b be a runner's finishing time relative to their time goal and their best marathon time, respectively. We estimate a multiple reference point model that is the sum of satisfaction

relative to the time goal,

$$S_g(x_g) = \begin{cases} n_{1g}x_g + n_{2g}x_g^2 , & x_g < 0 \\ p_{1g}x_g + p_{2g}x_g^2 + J_g, & x_g \geq 0 \end{cases},$$

and satisfaction relative to best marathon time,

$$S_b(x_b) = \begin{cases} n_{1b}x_b + n_{2b}x_b^2 , & x_b < 0 \\ p_{1b}x_b + p_{2b}x_b^2 + J_b, & x_b \geq 0 \end{cases}.$$

Table 4 shows the estimates of this model, and Figure 5 displays the separate marginal effects of performance relative to the time goal and relative to best marathon time, generated by holding performance relative to the other reference point at zero. Although the multiple reference point model provides a significantly better fit than the model incorporating only the time goal as a reference point (likelihood ratio test, $\chi^2(5) = 219.14$, $p < .001$), this comparison mostly impacts overall satisfaction by producing an additional jump in satisfaction when performance exceeds a runner's best marathon time, $\chi^2(1) = 13.19$, $p < .001$. Thus, we find evidence that time goal and best marathon time simultaneously act as reference points in shaping satisfaction. We find nearly identical results if we use last marathon time in place of best marathon time (see Electronic supplementary material, Section A.3.5).

	Predictor	Time Goal		Best Time	
Linear	losses	$(\hat{n}_{1g})$	18.59***	$(\hat{n}_{1b})$	−1.64*
	gains	$(\hat{p}_{1g})$	34.09**	$(\hat{p}_{1b})$	1.02
Quadratic	losses	$(\hat{n}_{2g})$	19.06***	$(\hat{n}_{2b})$	−1.30
	gains	$(\hat{p}_{2g})$	−306.83	$(\hat{p}_{2b})$	−1.95
	Jump	$(\hat{J}_g)$	0.74***	$(\hat{J}_b)$	0.80***

Observations	1015
degrees of freedom	25
-2 log likelihood	3071.98

Table 4: Multiple reference point analysis. Ordered logit regression on quadratic polynomial relating satisfaction to normalized relative performance, defined relative to both time goal and a runner's best previous marathon time. The statistical significance of coefficients is indicated by: *** ($p < .01$), ** ($p < .05$), and * ($p < .10$).
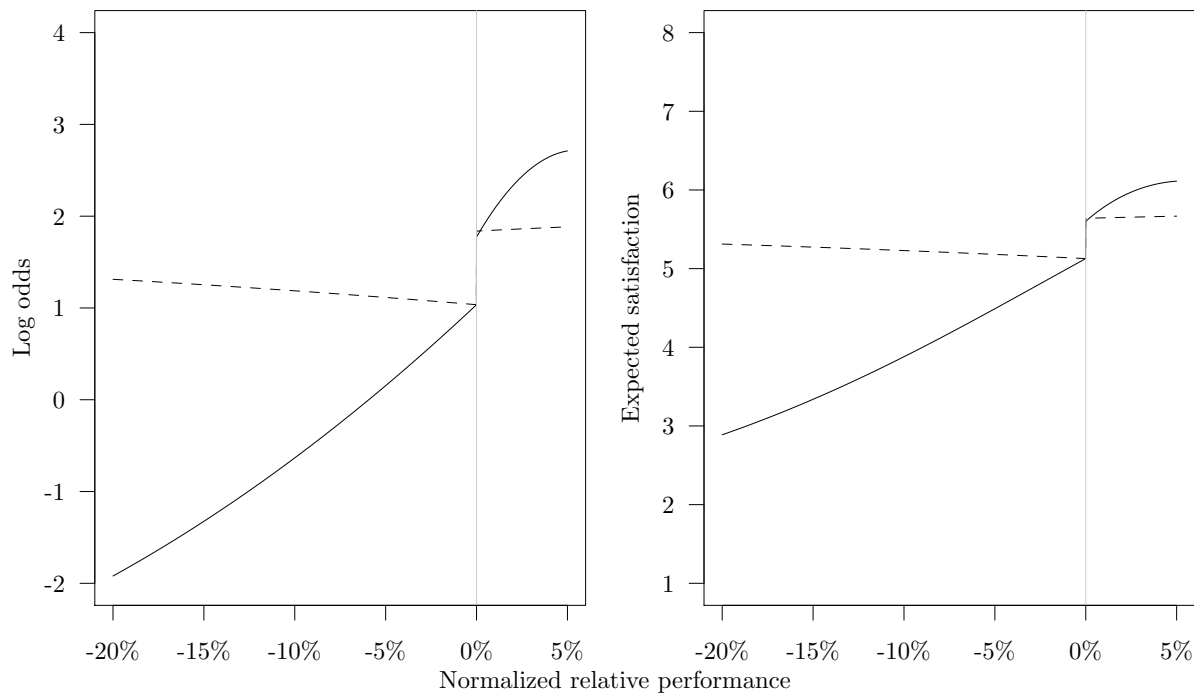
Figure 5: Marginal effects of normalized performance relative to time goal (solid), and relative to best marathon time (dashed), on satisfaction. The left panel plots the log odds of meeting or exceeding the midpoint of our satisfaction scale (4), while the right panel presents expected satisfaction on the original ordinal (1-7) scale. For each marginal effect plot, performance relative to the other reference point was held at zero.

6   Alternative Explanations

We have suggested that our results provide evidence that goals act as reference points and therefore that satisfaction exhibits loss aversion and diminishing sensitivity relative to the time goal reference point. In this section, we consider three alternative explanations for our finding.

6.1   Is this pattern driven by participants changing their goal?

All of our analyses have used post-marathon goal measures to determine whether a participant exceeds or surpasses a goal. As a result, it is possible that our pattern of results is driven by a group of individuals who view achieving the goal as sufficiently important that they modify their goal after the race to appear as if they achieved their goal. In addition, 39.0% of runners who fell short of their pre-marathon goal increased their reported goal following the marathon, compared to 6.0% of the runners who bested their pre-marathon goal ($\chi^2(1, 836) = 25.44$, $p < .001$). Of

27

course, such a shift could occur for non-psychological reasons (such as an unfavorable change in the weather or an injury onset) or for psychological reasons (such as the desire to self-enhance).

To test for the possibility that changes in goals is driving the pattern in Figure 4, we examined all participants in the goal-asked condition who provided goals prior to and after the marathon ($n = 835$) and fit a piecewise quadratic polynomial to normalized relative performance defined by their pre-marathon goal as well as their post-marathon goal. The results of this analysis show a significant jump when we use pre-marathon goals ($\hat{J} = 1.13$, $p < .001$), as well as post-marathon goal ($\hat{J} = 1.08$, $p < .001$), suggesting that our main results are not driven by shifts in goals. Details of this analysis are found in the Electronic supplementary material, Section A.3.6.

6.2   Boston Marathon qualifying

We have interpreted the discontinuity in satisfaction at a time goal as being psychological in nature. However, the pattern could reflect extrinsic benefits, most notably qualifying for the Boston Marathon. The Boston Marathon is America's oldest marathon and requires that the bulk of its participants qualify for the race by meeting a standard that differs for age and gender. To examine whether our effect is driven by this extrinsic benefit, we include a dummy variable that captures whether a participant's time goal is within one minute on either side of the Boston Marathon qualifying time for their gender and age group ($n = 150$). We then repeat our analysis by including the Boston Marathon qualifying dummy variable as well as a term that interacts our jump parameter with the Boston Marathon qualifying dummy variable. The jump remains positive and significant ($\hat{J} = 0.92$, $p < .001$), suggesting that our pattern of satisfaction is not driven by the extrinsic benefit of qualifying for the Boston Marathon.

6.3   Unobserved Heterogeneity

Because we did not randomly assign goals to marathon participants, it is possible that marathoners who surpass their goals differ from runners who fall short of their goals in respects other than their performance. A similar critique applies to Boyce et al. (2013). In other words, a value function that appears to exhibit reference dependence might capture a mixture of agents, none of whom exhibit the characteristic Prospect Theory value function. To illustrate, we sketch a two-type model and show how this model could produce a pattern that resembles Figure 4 (with the exception of

diminishing sensitivity).

Suppose that there are two types of runners, "ambitious" (denoted $\Theta = A$) and "unambitious" (denoted $\Theta = U$). Ambitious runners set aggressive goals and achieve these goals very infrequently. Unambitious runners establish modest goals and mostly achieve these goals. If satisfaction for ambitious runners is linear and monotonically increasing in performance relative to the goal, and satisfaction for unambitious runners is flat and high (above 6 on the 1-to-7 Likert scale), a mixture of these two types could easily produce what appears to be loss aversion and/or a jump in aggregate, even though both types have a linear satisfaction function.

We examine whether our data are consistent with this two-type model by exploiting a natural experiment in our data. Our data sample includes 15 marathons over 3 years. These marathons differ considerably in the overall rate in which marathoners in our sample achieve their goals. Only 4.0% of Chicago Marathon runners in 2007 met their goal, compared to 45.0% of runners in the 2009 Chicago Marathon. Not surprisingly, the rates of goal attainment are driven mostly by weather. The high temperature for the Chicago Marathon in 2007 was 93 degrees, while the high temperature in 2009 was 43 degrees (see Electronic supplementary material, Table A.6). Across marathons, the correlation between the rate of goal attainment and the day's high temperature is $r = -.78$.

If the two-type model holds and the exogenous shock of weather does not influence the proportion of types $\Theta = A$ and $\Theta = U$, only the likelihood that ambitious runners achieve their goals, then we should see more loss aversion and larger jumps for low success rate marathons and less loss aversion and smaller jumps for high success rate marathons. To test this implication, we refit our quadratic model and include the high temperature for each marathon and a term that interacts high temperature and whether a participant met their time goal. We also include gender but omit dummy variables for each marathon, because variation across marathons is captured by the weather variable. Critically and contrary to the two-type model, the interaction term is not significant ($\chi^2(1) = 0.35$, $p = 0.552$), and when the interaction term is included, the jump is significant even for the 2007 Chicago Marathon ($\hat{J} = .79$, $p = .031$). We obtain nearly identical results using low and average temperatures for each marathon.

7 Discussion

Investigations of reference-dependent preferences have typically involved two domains: risky decision making and the endowment effect. While these domains differ in critical ways, they share important commonalities – traditionally, both are investigations of decision making around a status quo reference point, either current wealth level or state of ownership of a good.

Although taking the status quo as a reference point is sensible in these studies, for many decisions, other standards may be more relevant (e.g., Barberis 2013; Baucells et al. 2011). In addition, the use of choice and pricing as dependent measures generates insights about decision utility but is typically agnostic about the welfare implications of those decisions. Our study examines a domain distinct from most prior studies. We study how goals influence satisfaction, and thus our investigation departs from most research on reference dependence both in using a non-status quo reference point and in enlisting satisfaction as a dependent measure. Specifically, we show that satisfaction with performance for marathon runners is reference-dependent, with satisfaction exhibiting a jump at the reference point as well as classical loss aversion, an asymmetry in the steepness of loss and gain slopes. In addition, satisfaction is influenced by multiple reference points – although a marathoner's time goal is the primary reference point, best and last marathon times serve as secondary benchmarks. While earlier studies have also found evidence that multiple reference points can operate simultaneously (Ordóñez et al. 2000), our study does so in a naturalistic manner without specifically drawing a participant's attention to possible alternative reference points. Finally, by documenting loss aversion in both predicted and experienced utility, we show that loss aversion is not a forecasting error.

We conclude by discussing the implications of our research for three broad questions about reference points. First, what is the relationship between decision utility, predicted utility, and experienced utility? Second, does a jump in utility at the reference point generalize beyond marathon runners and satisfaction? Third, what psychological processes govern the formation and adaptation of reference points?

7.1 Decision Utility, Predicted Utility, and Experienced Utility

Prospect theory and other models of reference-dependent utility (e.g., Kőszegi and Rabin 2006, 2007) are traditionally regarded as accounts of *decision utility*, i.e., specifications of how probabilities and outcomes are evaluated and combined to explain and predict decisions. Kahneman, Wakker, and Sarin (1997) contrast decision utility with two other notions of utility: *experienced utility*, the hedonic experience associated with an outcome, and *predicted utility*, an individual's forecast of her experienced utility. A correspondence between these three types of utility is often implicitly assumed because "choices provide all necessary information about the utility of outcomes because rational agents who wish to do so will optimize their hedonic experience" (Kahneman et al. 1997, p. 375).

Studies have documented a deviation between decision and experienced utility (e.g., Scitovsky 1976; Kahneman, Frederickson, Schreiber, and Redelmeier 1993) and between predicted and experienced utility (e.g., Kahneman and Snell 1990; Loewenstein and Adler 1995). However, most of these studies have not investigated the relationship between decision, predicted, and experienced utility in settings involving reference points. One exception is Kermer et al. (2006), who found loss aversion in predicted utility but not in experienced utility. These results led the authors to conclude in the title of their paper that "loss aversion is an affective forecasting error." Some recent studies, however, have cast doubt on this conclusion by documenting loss aversion in experienced utility. In a large-scale study involving 49,393 individuals from Germany and the United Kingdom, Boyce et al. (2013) found that a loss of income had a more substantial effect on subjective well-being than an equivalent income gain. Ockenfels, Sliwka, and Werner (2015) documented a similar pattern in a study of German personnel data, demonstrating loss aversion in job satisfaction for deviations around a fair bonus payment. Finally, Carter and McBride (2013) fit satisfaction scores elicited from an experimental "Matching Pennies" game and uncovered an S-shaped satisfaction with loss aversion for large (but not small) losses.

Our study provides some additional insights into the important question of how decision, predicted, and experienced utility are related in settings involving possible gains and losses relative to a reference point. We find loss aversion in both predicted utility (Figure 2) and experienced utility (Figure 4). It is important to note that our findings merely document a qualitative correspondence

between predicted and experienced utility when goals serve as a reference point. We do not make any quantitative claims on whether the in the degree of loss aversion differs for prediction relative to experience. Measurement of loss aversion is undoubtedly highly sensitive to the salience of the reference point. Therefore, methods for eliciting predicted versus experienced satisfaction may very well not be psychologically neutral in drawing attention to a reference point. As a result, we refrain from making statements about whether there is more or less loss aversion in predicted or actual satisfaction, merely noting that both types of judgments exhibit loss aversion. In addition, because our study only examines aggregate patterns of satisfaction, it is silent about the correspondence between predicted and experienced utility at the individual level.[21]

Although our study does not speak directly to decision utility when goals serve as reference points, Allen, Dechow, Pope, and Wu (2017) document a pattern of marathon finishing times consistent with a value function with a jump, classical loss aversion, and diminishing sensitivity. They find significant bunching near round-number finishing times, such as 3:00 or 3:30. For example, 50.0% more runners complete a marathon in 2:59:xx than in 3:00:xx. The archival data analyzed in Allen et al. and the survey data presented in this paper collectively indicate that utility, whether that utility is decision, predicted, or experienced, exhibits loss aversion and diminishing sensitivity when goals serve as reference points.

Of course, this paper leaves unanswered the more general question of the relationship between decision, predicted, and experienced utility for settings with non-goal reference points. The question of whether goals operate differently from other reference points, status quo or otherwise, is beyond the scope of this study. We hope that our study serves as an inspiration and possible template for important investigations of this sort.

7.2 Jumps at the Reference Point

The discontinuity or jump in satisfaction at the goal documented in this investigation raises the natural question of whether our observed pattern of experienced utility extends to a gambling

---

[21]Both of these questions, while important, pose empirical challenges. For example, focalism might produce differences in prediction and experience (Wilson et al. 2000). In prediction, marathon runners could focus on how falling short or exceeding a goal influences their satisfaction and ignore other factors (such as the weather, injuries, etc.) that also likely affect satisfaction. Focalism, therefore, could result in differences in loss aversion, as predicted by Kermer et al. (2006) or merely a more compressed relationship between satisfaction and relative performance in experience. In addition, asking a participant for multiple predictions of satisfaction might highlight the difference between gains and losses (e.g., McGraw et al. 2010).

context or other settings in which Prospect Theory has been evoked. On the one hand, there are psychological reasons that our marathon setting may differ from investigations of risky choice such as Abdellaoui et al. (2007). The qualitative difference between barely surpassing and just falling short of a performance goal may be different because goal pursuit is in large part driven by the desire to achieve success and avoid failure (e.g., Atkinson 1957). The notion that goals may produce a more dichotomous categorization of outcomes than other types of reference points suggests that the value function around a goal may have a distinct shape.

There are other reasons, however, to surmise that a jump at the reference point is not restricted to goal pursuit. Studies of risky decision making have documented patterns consistent with a discontinuous value function. Schneider and Lopes (1986) found risk-seeking in losses involving a chance at breaking even, but no risk-seeking when loss gambles were displaced from zero, a pattern consistent with a value function that is linear in losses with a jump at the reference point of 0. In addition, Payne (2005) and, to a lesser extent, Diecidue, Levy, and van de Ven (2015), provide some evidence in risky choice for a jump at 0.

One possible reason for the limited evidence for jumps at the reference point is that statistical identification of a jump is difficult, even though loss aversion can be decomposed conceptually into loss aversion at and away from the reference point. In our analyses, although the jump at the reference point is robust to various model specifications, we do observe a tradeoff between the size of the jump and the magnitude of the slope in the loss domain. In addition, parametric models may be unable to distinguish between a continuous value function that is very steep near the reference point and a discontinuous model with a jump at the reference point. Identification requires specially-designed stimuli with a relatively dense set of outcomes located near the reference point. Interestingly, reference-dependent preferences in decision utility generate a bunching of finishing times around a goal (e.g., Allen et al. 2017). In turn, identification of reference-dependent experienced utility in this study is facilitated by the saturation of relative performance around 0.

Although we find evidence for a jump in satisfaction at the reference point, we suspect that the value function around a goal might not be well-behaved. As an analogy, recall that Kahneman and Tversky (1979) originally proposed a probability weighting function that was discontinuous at 0 and 1. Kahneman and Tversky suggested that "highly unlikely events are either ignored or overweighted...Consequently, π is not well-defined near the end-points." (p. 283). A similar

psychology may apply to performance near a goal. Falling a bit short of a goal will sometimes be rounded up and coded as a "success." In other cases, the same outcome will be viewed as "falling short" and hence be seen as a substantial setback. The implication then is poorly-specified behavior around a reference point, with some losses assimilated to the reference point and other losses standing out as disappointing.

7.3 Reference Point Formation

Earlier, we discussed some of the challenges with identifying the appropriate reference point for a particular application. In investigations of asset pricing, for example, the reference point may be the purchase price, the current price, the historical high price, or some value determined by an investor's expectations. The difficulty in adjudicating between a set of plausible reference points reflects in part an impoverished understanding of what psychological processes drive reference point formation and adaptation (for some exceptions, see Arkes et al. 2008, Baillon et al. 2017, and Baucells et al. 2011). While a general account of reference-point formation is lacking, we propose that motivational processes can be implicated in the selection of reference points for achievement-oriented activities. This study and Allen et al.'s (2017) related work provide insight into how goals operate once they have been formed. Marathoners often choose round number goals (49.1% of our sample chose a half-hour or hour-goal). Goals, whether or not a round number, serve as reference points, influencing decision utility through allocation of effort in training and on race day, as well as experienced utility, how marathoners judge their performance.

Our study is, however, silent about how goals are determined or how they might change over time. Decision makers who self-set goals face a dilemma. Although higher goals generally lead to higher performance (Locke and Latham 2006), those who perform better, paradoxically, often feel worse because their performance falls below those lofty goals (Mento et al. 1992), a direct implication of a "goals as reference points" account. Goal formation processes therefore must balance the self-regulatory value of difficult goals with a desire to avoid disappointment with one's performance. Recent theoretical models in economics such as Hsiaw (2013) have tried to resolve this paradox by studying goals within a dual self framework. In addition, some recent empirical work by Clark et al. (2017) and Corgnet et al. (2015) provides some fresh data on the relationship between goal selection, monetary rewards, and performance.

We hope that a better understanding of the relationship between goals and satisfaction, as put forth in this paper, provides a starting point for determining how goals are formed and updated over time. In other settings, reference points might be determined by other psychological processes, including basic cognitive processes such as attention, memory, or categorization, as well as hedonic and motivational considerations. If the results shown here are general, both decision making and satisfaction will be governed by a calculus in which outcomes that fall short of or exceed the reference point, however determined, are viewed as qualitatively different.

References

Abdellaoui, M. (2000). Parameter-free elicitation of utility and probability weighting functions. *Management Science*, 46(11), 1497–1512.

Abdellaoui, M., Bleichrodt, H., & Paraschiv, C. (2007). Loss aversion under prospect theory: A parameter-free measurement. *Management Science*, 53(10), 1659–1674.

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

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

Arkes, H.R., Hirshleifer, D., Jiang, D., & Lim, S. (2008). Reference point adaptation: Tests in the domain of security trading. *Organizational Behavior and Human Decision Processes*, 105(1), 67–81.

Atkinson, J.W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review*, 64(6p1), 359.

Austen, I. (2001). You clocked what? For marathon runners, It's gun vs. chip. *The New York Times*, 1 November 2001.

Baillon, A., Bleichrodt, H., & Spinu, V. (2017). Searching for the reference point. Working paper.

Barberis, N. & Xiong, W. (2009). What drives the disposition effect? An analysis of a long-standing preference-based explanation. *Journal of Finance*, 64(2), 751–784.

Barberis, N.C. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1), 173–195.

Bartling, B., Brandes, L., & Schunk, D. (2015). Expectations as reference points: Field evidence from professional soccer. *Management Science*, 61(11), 2446–2461.

Baucells, M., Weber, M., & Welfens, F. (2011). Reference-point formation and updating. *Management Science*, 57(3), 506–519.

Benartzi, S. & Thaler, R.H. (1995). Myopic loss aversion and the equity premium puzzle. *Quarterly Journal of Economics*, 110(1), 73–92.

Bentham, J. (1789). *An Introduction to the Principles of Morals and Legislation.* Clarendon Press, Oxford, UK.

Booij, A.S. & Van de Kuilen, G. (2009). A parameter-free analysis of the utility of money for the general population under prospect theory. *Journal of Economic Psychology*, 30(4), 651–666.

Boyce, C.J., Wood, A.M., Banks, J., Clark, A.E., & Brown, G.D. (2013). Money, well-being, and loss aversion does an income loss have a greater effect on well-being than an equivalent income gain? *Psychological Science*, 24(12), 2557–2562.

Briesch, R.A., Krishnamurthi, L., Mazumdar, T., & Raj, S.P. (1997). A comparative analysis of reference price models. *Journal of Consumer Research*, 24(2), 202–214.

Burdina, M., Hiller, R.S., & Metz, N.E. (2017). Goal attainability and performance: Evidence from boston marathon qualifying standards. *Journal of Economic Psychology*, 58(1), 77–88.

Camerer, C. (2005). Three cheers – psychological, theoretical, empirical – for loss aversion. *Journal of Marketing Research*, 42(2), 129–133.

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

Carter, S. & McBride, M. (2013). Experienced utility versus decision utility: Putting the 's' in satisfaction. *Journal of Socio-Economics*, 42, 13–23.

Clark, D., Gill, D., Prowse, V., & Rush, M. (2017). Using goals to motivate college students: Theory and evidence from field experiments. IZA Discussion Papers, No. 10283.

Corgnet, B., Gómez-Miñambres, J., & Hernán-Gonzales, R. (2015). Goal setting and monetary incentives: When large stakes are not enough. *Management Science*, 61(12), 2926–2944.

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

Curtin, R., Presser, S., & Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly*, 64(4), 413–428.

Diecidue, E., Levy, M., & van de Ven, J. (2015). No aspiration to win? An experimental test of the aspiration level model. *Journal of Risk and Uncertainty*, 51(3), 245–266.

Diecidue, E. & Van de Ven, J. (2008). Aspiration level, probability of success and failure, and expected utility. *International Economic Review*, 49(2), 683–700.

Draper, N.R. & Smith, H. (1966). *Applied regression analysis*. Wiley, New York, NY.

Ericson, K.M.M. & Fuster, A. (2011). Expectations as endowments: Evidence on reference-dependent preferences from exchange and valuation experiments. *Quarterly Journal of Economics*, 126(4), 1879–1907.

Farber, H.S. (2005). Is tomorrow another day? The labor supply of New York City cab drivers. *Journal of Political Economy*, 113(1).

Farber, H.S. (2008). Reference-dependent preferences and labor supply: The case of New York City taxi drivers. *American Economic Review*, 98(3), 1069–1082.

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

Fehr, E., Hart, O.D., & Zehnder, C. (2011). How do informal agreements and renegotiation shape contractual reference points? National Bureau of Economic Research Working Paper No. 17545.

Frey, B. & Stutzer, A. (2002). The economics of happiness. *World Economics*, 3(1), 25–41.

Fryer Jr, R.G., Levitt, S.D., List, J., & Sadoff, S. (2012). Enhancing the efficacy of teacher incentives through loss aversion: a field experiment. Natonal Bureau of Economic Research Working Paper 18237.

Gächter, S., Johnson, E.J., & Herrmann, A. (2007). Individual-level loss aversion in riskless and risky choices, IZA Discussion Paper no. 2961. Technical report.

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

Gilbert, D.T., Pinel, E.C., Wilson, T.D., Blumberg, S.J., & Wheatley, T.P. (1998). Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, 75(3), 617–638.

Goldstein, E.R. (2011). The anatomy of influence. *Chronicle of Higher Education*, 58(13), B6–B10.

Green, P.J. & Silverman, B.W. (1993). *Nonparametric regression and generalized linear models: A roughness penalty approach.* CRC Press.

Harrell, F.E. (2010). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* Springer Science & Business Media, New York, NY.

Hart, O. & Moore, J. (2008). Contracts as reference points. *The Quarterly Journal of Economics*, 123(1), 1–48.

Hastie, R. & Dawes, R.M. (2001). *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making.* SAGE, Thousand Oaks, CA.

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

Heath, C., Larrick, R.P., & Wu, G. (1999). Goals as reference points. *Cognitive Psychology*, 38(1), 79–109.

Higdon, H. (2011). *Marathon: The Ultimate Training Guide.* Rodale, Emmaus, PA.

Ho, T.H. & Zhang, J. (2008). Designing pricing contracts for boundedly rational customers: Does the framing of the fixed fee matter? *Management Science*, 54(4), 686–700.

Hsiaw, A. (2013). Goal-setting and self-control. *Journal of Economic Theory*, 148(2), 601–626.

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

Kahneman, D. (1999). Objective happiness. In D.E. Kahneman, E.E. Diener, & N.E. Schwarz, editors, *Well-being: The Foundations of Hedonic Psychology*, pages 3–25. (Russell Sage Foundation, New York, NY).

Kahneman, D., Fredrickson, B.L., Schreiber, C.A., & Redelmeier, D.A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, 4(6), 401–405.

Kahneman, D., Knetsch, J.L., & Thaler, R.H. (1990). Experimental tests of the endowment effect and the Coase Theorem. *Journal of Political Economy*, 98(6), 1325–1348.

Kahneman, D. & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, 39(1), 17–31.

Kahneman, D. & Snell, J.S. (1990). *Predicting Utility*, pages 295–310. University of Chicago Press, Chicago.

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

Kahneman, D., Wakker, P.P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics*, 112(2), 375–406.

Kass, R.E. & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.

Kermer, D.A., Driver-Linn, E., Wilson, T.D., & Gilbert, D.T. (2006). Loss aversion is an affective forecasting error. *Psychological Science*, 17(8), 649–653.

Köbberling, V. & Wakker, P.P. (2005). An index of loss aversion. *Journal of Economic Theory*, 122(1), 119–131.

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

Kőszegi, B. & Rabin, M. (2007). Reference-dependent risk attitudes. *American Economic Review*, 97(4), 1047–1073.

Krawczyk, M. & Wilamowski, M. (2017). Are we all overconfident in the long run? evidence from one million marathon participants. *Journal of Behavioral Decision Making*, 30(3), 719–830.

Larsen, J.T., McGraw, A.P., Mellers, B.A., & Cacioppo, J.T. (2004). The agony of victory and thrill of defeat: Mixed emotional reactions to disappointing wins and relieving losses. *Psychological Science*, 15(5), 325–330.

List, J.A. (2003). Does market experience eliminate market anomalies? *Quarterly Journal of Economics*, 118(1), 41–71.

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

Loewenstein, G. & Adler, D. (1995). A bias in the prediction of tastes. *Economic Journal*, 105(431), 929–937.

Long, J.S. & Freese, J. (2014). *Regression models for categorical dependent variables using Stata*. Stata Press, College Station, TX.

Lopes, L.L. (1987). Between hope and fear: The psychology of risk. *Advances in Experimental Social Psychology*, 20, 255–295.

Lopes, L.L. & Oden, G.C. (1999). The role of aspiration level in risky choice: A comparison of cumulative prospect theory and SP/A theory. *Journal of Mathematical Psychology*, 43(2), 286–313.

March, J.G. & Shapira, Z. (1992). Variable risk preferences and the focus of attention. *Psychological Review*, 99(1), 172–183.

Mas, A. (2006). Pay, reference points, and police performance. *Quarterly Journal of Economics*, 121(3), 783–821.

McGraw, A.P., Larsen, J.T., Kahneman, D., & Schkade, D. (2010). Comparing gains and losses. *Psychological Science*, 21(10), 1438–1445.

Mento, A.J., Locke, E.A., & Klein, H.J. (1992). Relationship of goal level to valence and instrumentality. *Journal of Applied Psychology*, 77(4), 395–405.

Ockenfels, A., Sliwka, D., & Werner, P. (2014). Bonus payments and reference point violations. *Management Science*, 61(7), 1496–1513.

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

Oettinger, G.S. (1999). An empirical analysis of the daily labor supply of stadium venors. *Journal of Political Economy*, 107(2), 360–392.

Ordóñez, L.D. (1998). The effect of correlation between price and quality on consumer choice. *Organizational Behavior and Human Decision Processes*, 75(3), 258–273.

Ordóñez, L.D., Connolly, T., & Coughlan, R. (2000). Multiple reference points in satisfaction and fairness assessment. *Journal of Behavioral Decision Making*, 13(3), 329–344.

Payne, J.W. (2005). It is whether you win or lose: The importance of the overall probabilities of winning or losing in risky choice. *Journal of Risk and Uncertainty*, 30(1), 5–19.

Pope, D. & Simonsohn, U. (2011). Round numbers as goals evidence from baseball, SAT takers, and the lab. *Psychological Science*, 22(1), 71–79.

Pope, D.G. & Schweitzer, M.E. (2011). Is Tiger Woods loss averse? Persistent bias in the face of experience, competition, and high stakes. *American Economic Review*, 101(1), 129–157.

Post, T. & van den Assem, M.J. (2008). Deal or no deal? Decision making under risk in a large-payoff game show. *American Economic Review*, 98(1), 38–71.

Prelec, D. (1998). The probability weighting function. *Econometrica*, 66(3), 497–528.

Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica*, 68(5), 1281–1292.

Sackett, A.M., Wu, G., White, R.J., & Markle, A.B. (2015). Harnessing optimism: How eliciting goals improves performance. Working paper.

Samuelson, W. & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1), 7–59.

Schmidt, U. & Zank, H. (2005). What is loss aversion? *Journal of Risk and Uncertainty*, 30(2), 157–167.

Schneider, S.L. & Lopes, L.L. (1986). Reflection in preferences under risk: Who and when may suggest why. *Journal of Experimental Psychology: Human Perception and Performance*, 12(4), 535–548.

Schwarz, N. & Strack, F. (1999). *Reports of subjective well-being: Judgmental processes and their methodological implications*, pages 61–84. Russell Sage Foundation, New York, NY.

Scitovsky, T. (1976). *The Joyless Economy: An inquiry into human satisfaction and consumer dissatisfaction.* Oxford University Press, Oxford, England.

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

Sullivan, K. & Kida, T. (1995). The effect of multiple reference points and prior gains and losses on managers' risky decision making. *Organizational Behavior and Human Decision Processes*, 64(1), 76–83.

Tovar, P. (2009). The effects of loss aversion on trade policy: Theory and evidence. *Journal of International Economics*, 78(1), 154–167.

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

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

Wang, X.T. & Johnson, J.G. (2012). A tri-reference point theory of decision making under risk. *Journal of Experimental Psychology: General*, 141(4), 743–756.

Wang, Y. (2011). *Smoothing Splines: Methods and Applications.* CRC Monographs on Statistics & Applied Probability. Chapman and Hall.

Weingarten, E., Bhatia, S., & Mellers, B. (2016). Multiple goals as reference points. Working paper.

Wilson, T.D. & Gilbert, D.T. (2003). Affective forecasting. *Advances in Experimental Social Psychology*, 35, 345–411.

Wilson, T.D., Meyers, J., & Gilbert, D.T. (2003). "How happy was I, anyway?" A retrospective impact bias. *Social Cognition*, 21(6), 421–446.

Wilson, T.D., Wheatley, T., Meyers, J.M., Gilbert, D.T., & Axsom, D. (2000). Focalism: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, 78(5), 821836.

Winer, R.S. (1986). A reference price model of brand choice for frequently purchased products. *Journal of Consumer Research*, 13(2), 250–256.

Wu, G. & Gonzalez, R. (1996). Curvature of the probability weighting function. *Management science*, 42(12), 1676–1690.

Wu, G. & Markle, A.B. (2008). An empirical test of gain-loss separability in prospect theory. *Management Science*, 54(7), 1322–1335.

Zeisberger, S., Langer, T., & Weber, M. (2012). Why does myopia decrease the willingness to invest? Is it myopic loss aversion or myopic loss probability aversion? *Theory and Decision*, 72(1), 35–50.