The Structure of Health Incentives: Evidence from a Field Experiment

Mariana Carrera[a], Heather Royer[b], Mark Stehr[c], Justin Sydnor[d]

Contact mariana.carrera@montana.edu (MC); hnroye@ucsb.edu (HR); stehr@drexel.edu (MS); justin.sydnor@wisc.edu (JS)

[a]College of Agriculture, Montana State University, Bozeman Montana (59717); [b]Department of Economics, University of California at Santa Barbara, Santa Barbara, California (93106), NBER & IZA; [c]LeBow College of Business, Drexel University, Philadelphia, Pennsylvania (19104); [d]Wisconsin School of Business, University of Wisconsin, Madison, WI, (53706), & NBER.

A growing number of firms use incentive programs to encourage healthy behaviors, but there is little evidence about how such incentives should be structured over time. We explore this issue using a large field experiment that incentivized employees of a Fortune 500 company to use their workplace gym. We compare the effectiveness of a treatment with constant incentives over 8-weeks to two treatments that varied incentives over time. One variable treatment featured front-loaded incentives, which could, in theory, help procrastinators overcome startup costs to joining an incentive program. We find, however, that the front-loaded incentive did not increase participation on the extensive margin relative to the constant incentive and was less effective in sustaining exercise over time. The second variable incentive was designed to leverage short-term habit-formation by turning incentives on and off over a longer period of time. This novel sporadic incentive showed slightly stronger effects than the constant incentive. We discuss how the nature of habit formation processes affects the relative benefits of consistent versus periodic incentives.

Keywords: incentives, health behaviors, field experiment, exercise, wellness program, gym, procrastination, present bias, habit formation

This study was funded by a grant from the Robert Wood Johnson Foundation through the Applying Behavioral Economics to Perplexing Problems in Health and Health Care initiative via the Center for Health Incentives and Behavioral Economics at the University of Pennsylvania. Stehr thanks the LeBow College of Business for support from the Dean's Research Fellowship. We appreciate the work of a dedicated team of research assistants, including Unnati Amin, Steven Cramer, Nicolas Erturkuner, Jordan Hsieh, Chang Lee, Chenyuan Liu, Ben Marks, Sarah Mattson, Leah Neustadt, Charlie Nusbaum, Jacob Rabin, Alan Thomas, and Jove Tse. We also thank Gary Charness, Leslie John, and Katherine Milkman for helpful comments and suggestions.

Firms are increasingly using targeted financial incentives to encourage workers to improve their health behaviors. According to the Kaiser Family Foundation Employer Health Benefits Survey, 81% of large firms offered a wellness program in 2015 with 38% of these programs offering their employees monetary incentives for participation.[1] The Affordable Care Act also encourages the use of health incentives by allowing up to 30%of the cost of health care coverage to be allocated to wellness program rewards.[2]

A burgeoning literature explores how people respond to financial incentives to quit smoking, lose weight, eat healthier or exercise more (Finkelstein et al, 2007; Volpp et al., 2008; Charness and Gneezy, 2009; Volpp et al., 2009; Babcock and Hartman, 2010; Gneezy, Meier, and Rey-Biel, 2011; John et al., 2011; Cawley and Price, 2013; Just and Price, 2013; Milkman, Minson and Volpp, 2013; Acland and Levy 2015; Babcock et al., 2015; Halpern et al., 2015; Royer, Stehr and Sydnor, 2015; Loewenstein, Price and Volpp, 2016; Patel et al 2016; Mochon et al., 2017, Carrera et al., 2018). These studies show that incentives can meaningfully affect behavior while they are in place and in a few cases have even found persistent effects once incentives are removed (Charness and Gneezy, 2009; Loewenstein, Price and Volpp, 2016; Hussam et al., 2017; Mochon et al., 2017).

Relatively little research has explored how, given a fixed budget, incentives should be structured to generate behavior change.[3] Many programs offer fixed payments, such as a payment for each gym visit or each day below a weight target, that create a constant incentive to change behavior over a period of time. Might programs that vary incentives over time be more effective? If the targeted health activity involves startup costs, the potential for habit formation, or both, then differently structured incentives may have different impacts. Charness and Gneezy (2009) were the first to highlight that incentives may be important for overcoming "initial resistance to commencing a beneficial regimen" in the face of startup costs and that incentives that generate enough instances of the activity may "move some people past the 'threshold'" of habit formation necessary to sustain behavior.

[1] Web link: http://kff.org/report-section/ehbs-2015-summary-of-findings/

[2] See Federal Register Doc No: 2013-12916.

[3] Two notable exceptions are Haisley et al. (2012) who find that a lottery design is more effective at encouraging people to complete a health-risk assessment than a fixed payment with the same expected value, and Patel et al (2016) who do not see increased effects of a lottery-framed incentive but do find a loss-framed incentive is more effective than an equivalent gain-framed incentive.

To build intuition about incentive design, we develop a simple model of daily health-behavior decisions for an agent with naïve present bias (O'Donoghue and Rabin, 1999, DellaVigna and Malmendier, 2006) that incorporates the potential for both one-time startup costs and habit formation (Becker and Murphy, 1988). The model illustrates that contingent incentives can help present-biased agents stop procrastinating, overcome startup costs, and invest in beneficial habit formation. The model also highlights that different incentive structures can interact with these considerations in important ways. In particular, if there are one-time startup costs on the extensive margin (e.g., joining the gym), some people who intend to begin the activity may continually procrastinate and a front-loaded incentive may be helpful for breaking that procrastination.[4] However, there is a tradeoff because front-loaded incentives require smaller incentives later, which may be insufficient to sustain behavior.

We explore these issues using a randomized field experiment incentivizing exercise at the on-site company gym for 980 employees of a Fortune 500 company. The two primary treatment arms for the experiment contrast a front-loaded incentive ("kickstart") versus a consistent incentive ("constant"). These two treatments offered the same maximum possible earnings ($160) over 8-weeks but differed in how the earnings were allocated over time. The constant treatment offered $10/visit for up to two visits each week. In contrast, the kickstart treatment front-loaded incentives, offering $25/visit in the first two weeks followed by $5/visit in the remaining six weeks.[5] Employees of this company must join the fitness facility before using it, which creates an identifiable group (non-members) who face startup costs to using the company gym. For this group we can study the extensive-margin effects of the incentive program by looking at the fraction who join the gym.

Our primary finding is that the front-loaded program showed weaker overall effects than the constant incentive during the incentive period. We expected ex-ante that this kickstart program might be more effective among those who were not previously members of the gym due to their startup costs. We find, however, that for non-members, the join rates for those in the constant and kickstart programs were identical. Both were 10 percentage points higher than the join rate for

[4] In the savings literature, there is work on how the front-loading versus back-loading of taxes of retirement accounts affect savings (Thaler, 1994; Burman, Gale, and Weiner 2001; Beshears et al., 2017). Thaler (1994) highlighted that a Roth 401k might have negative effects because the incentive to invest is back-loaded. Empirical work that followed, though found little response to Roth 401k. The lack of response to differential 401k savings incentives, however, has largely been attributed to behavioral biases that are unlikely to extend to the exercise setting.

[5] As in the constant incentive, employees in the kickstart could earn the incentives for up to two visits each week.

those in our non-incentivized control condition. Our model predicts that the higher maximum daily incentive at the outset of a front-loaded scheme should matter for people who were intending to join the gym but were procrastinating each day due to the startup costs. The similar join rates suggest that the $10 incentive offered by the constant program may be sufficient for the procrastinating population to overcome their startup costs.

A secondary treatment offered to non-members helps to shed light on the failure of the kickstart incentive to induce a higher join rate than the constant incentive. The secondary treatment ("constant-short") used the same $10/visit structure as the constant incentive but lasted 4 weeks instead of 8 weeks. Since the peak incentive was the same as the baseline constant treatment, the model predicts that this shorter incentive would be equally effective at motivating joining among the population of procrastinators. However, the model also predicts that people who would *not* intend to join the gym in absence of incentives will base their participation decisions on the total anticipated earnings. Thus, the lower total earnings of the constant-short treatment should make it less attractive to those who were not planning to join. We find that the shorter treatment induced less than half as many people to join as the baseline treatments. Importantly, however, once the incentives were removed, the fraction continuing to exercise was only slightly smaller among the shorter 4-week program than among the 8-week program. This pattern suggests that the longer programs primarily added new members with low ex-ante value for this gym who failed to continue attending once incentives were removed.

The front-loaded kickstart program was also less effective than the constant incentive among the existing members of the gym. The model predicts that front loading could be beneficial for this population if larger daily incentives have a bigger impact on the probability of using the gym, especially if exercising can lead to rapid formation of habits. We find, however, that while both programs generated a significant increase in the fraction of existing members using the gym, the increase for the kickstart group during their initial two weeks with high incentives was only slightly elevated relative to the constant-incentive group. In contrast, the fraction using the gym was significantly lower for the kickstart group during the later period when their incentive was lower. The patterns suggest sharply diminishing returns to spot incentives for this population. These results are consistent with a concurrent working paper by Bachireddy et al. (2017), which finds that a constant incentive more effectively encouraged steps by participants with pedometers than continuously increasing or decreasing incentives with the same overall possible earnings.

We also find that the constant incentive, but not the kickstart, generated increases in gym use once the incentives were removed. In the eight weeks following the incentive program, the constant-incentive group retained on average about half of their elevated visit rates relative to the incentive period. However, the effect dissipated over time and they showed visit rates similar to the control group a few months after the incentive ended. So, while we see the effects of raised habit stocks on behavior consistent with classic theory on habit formation (Becker and Murphy, 1988), the results are not suggestive of habits crossing a "threshold" that generates a permanent new equilibrium of greater exercise frequency. Although Charness and Gneezy (2009) found stable persistence effects in their study of temporary exercise incentives, studies by Acland and Levy, (2015) and Royer et al., (2015) showed similar patterns of dissipating habits. It may be that it takes longer than 8 weeks to reach the threshold for a self-sustaining exercise habit among this working population. Or, as we discuss when presenting our model of health behaviors, habit formation may not always have a threshold character in which sufficient investment leads to a higher equilibrium (i.e. self-sustaining) level of the activity.

The possibility that incentives may boost habit stocks but not cause them to cross a self-sustaining threshold has implications for the design of incentives. Royer et al. (2015) pointed out that temporary boosts in habit stocks may favor programs that intersperse periods of incentives to promote behavior within un-incentivized periods by leveraging elevated habit stocks. We formalize that intuition and discuss model simulations comparing a constant incentive offered over a period of time versus the same incentive offered half as often (e.g., every other week) but for twice as long. In many cases where there is no self-sustaining threshold for higher habit stocks, the model predicts that the periodic incentive will produce moderately higher average exercise frequency over the longer period than the continuous incentive.

To test the potential benefits of periodic incentives, we tested a secondary treatment among the existing members of the company gym. This novel "extended-sporadic" design offered the same $10/visit structure for 8 weeks as the constant treatment, but participants were incentivized for 8 randomly-selected weeks out of 16 weeks. Consistent with the model predictions, the extended-sporadic treatment showed treatment effects over 16 weeks of between 60%-100% of the incentive-period effect for the constant incentive, compared to the 50% we would expect with no habit-formation effects. Considering average visits over 24 weeks from the start of the incentive programs, which includes 8 weeks following the extended-sporadic's incentive period, we find slightly (though not statistically significantly) higher effects for the extended-sporadic incentive than for the constant incentive. These results highlight that habit formation is often temporary and that intermittent incentives may be effective in leveraging temporarily elevated habit stocks in these situations.

2. Model

In this section, we introduce a simple model of daily exercise decisions for a naïve present-biased agent to build intuition for the tradeoffs involved in different incentive structures. We use a naïve version of present bias here because it is both consistent with some past evidence on exercise behavior (e.g., Della Vigna and Malmendier, 2006) and provides a simple framework for discussing important dynamics involving procrastination and underinvestment in establishing habits. We begin (Section 2.1) with a basic deterministic model comparing immediate costs and delayed benefits from exercising that provides a structure for considering the direct effect of a contingent incentive on the perceived benefit of exercise that day. We then enrich the model to consider how incentives interact with one-time "startup costs" that may generate procrastination (Section 2.2), simple "threshold" habit formation (Section 2.2), and finally a richer model incorporating stochastic exercise costs that allows for more continuous, i.e., non-threshold, responses to habit stocks (Section 2.3).

2.1 Model setup

Consider an individual who lives for T days and each day has the option to engage in an activity (e.g., exercise). The action has an immediate cost, c, and a delayed benefit, b, that arrives at some later date. The individual is naïve present-biased (O'Donoghue and Rabin, 1999), weighting all utility outcomes received in the future by a constant present-bias parameter $\beta \leq 1$ but wrongly believing that on future days she will not engage in this same discounting. We omit the standard exponential discounting term from the model for simplicity because none of the main insights change when exponential discounting is included along with present-bias. Anticipated remaining lifetime utility from the perspective of the current day t is given by:

$$U_t = a_t(\beta b - c) + \beta \sum_{j=t+1}^{T} \widehat{a_j}(b - c), \tag{1}$$

where a_t is an indicator that takes value 1 if the agent exercises that day and 0 otherwise. The individual believes she should and will exercise in the future if the undiscounted benefits of exercise are positive, $b - c > 0$. However, her beliefs about exercise in the future $\hat{a}_j$, may be wrong because she will actually exercise only if the discounted benefits exceed the current costs $\beta b - c > 0$. If $\beta b - c < 0 < b - c$, she will desire to exercise but fail to do so.

If this individual is offered an immediate incentive payment, p, contingent on exercising today, she will do so if $\beta b - c + p > 0$.[6] The effects of variable incentives in this simple situation will depend on the distribution of the underlying parameters in the population. For example, consider both a middle incentive level, p^m, as well as a higher p^h and lower incentive p^l that average to the same amount, i.e., $p^h + p^l = 2p^m$. People can be sorted into four groups by their response to these incentives: 1) "non-responders", who never exercise under any of these incentives ($\beta b - c + p^h < 0$), 2) "always responders", who exercise under all incentive levels ($\beta b - c + p^l > 0$), 3) "motivated by high", who exercise under the high incentive only ($\beta b - c + p^m < 0 < \beta b - c + p^h$), and 4) "motivated by middle", who exercise under the middle or high incentive but not the low ($\beta b - c + p^l < 0 < \beta b - c + p^m$). The relative effectiveness of the variable vs flat payment schemes will depend on the shares of people falling into group 3 (favors variable) versus those falling into group 4 (favors flat). In particular, if group 4 is much larger than group 3, there will be diminishing returns to incentive size and flat incentives will tend to be more effective than variable incentives.

2.2 One-time startup costs and threshold habit formation

We now add to this framework both the possibility of one-time startup costs and habit stocks. Both of these forces can favor using higher initial incentives to overcome procrastination by naïve agents. There are some differences in how these two forces interact with incentives, but

[6] We assume throughout that incentive payment benefits are received immediately. This is for simplicity only and does not affect the main modeling insights. Of course, the level of incentive needed to induce action would depend on the timing of incentive payment.

they share essential features. In fact, it is reasonable to think of initial periods of low habit stocks as another form of startup costs.

We first introduce one-time fixed "startup costs", f, to joining the gym (e.g., time, effort, psychological costs) that must be paid when exercising for the first time. Consider first a group we call the "procrastinators". For this group we assume that the net undiscounted daily benefits of exercise are positive, $b - c > 0$, and the agents will live sufficiently long that there is a clear benefit to paying the startup costs, $T(b - c) \gg f$. Without present bias, these individuals would clearly pay the startup costs in the first period and then exercise daily thereafter. However, as O'Donoghue and Rabin (1999) highlight, a naïve-present-biased agent may procrastinate indefinitely on doing actions with one-time up-front costs. Since a naïve agent believes she will be joining the gym the following day, whether she does so today only depends on her perceived utility of procrastinating for one day. She compares the net-value of exercise today $(\beta b - c)$ against the benefit of delaying the fixed cost by a day $(1 - \beta)f$ and joins today iff:

$$\beta b - c > (1 - \beta)f \qquad (2).$$

For high enough fixed costs or strong present bias, the agent will perpetually procrastinate on joining the gym, always believing,incorrectly, that she will join tomorrow.

Now consider how these "procrastinators" respond to different incentive structures. Recall these people believe they will join the gym and start exercising tomorrow regardless of the future flow of incentives. As such, only the spot incentive she faces today is relevant for whether she overcomes her procrastination. She will join the gym iff:

$$\beta b - c + p > (1 - \beta)f \qquad (3).$$

This means that for any incentive structure, among the "procrastinators", the *extensive margin* effect of the incentive on the number of people who join the gym will be dictated only by the size of the incentive on the day with the highest possible incentives. So, the extensive margin effects can be maximized for this group by front-loading the entire incentive budget on the first day of an incentive program. However, the *intensive margin* of the incentive scheme may be lower if incentives are front-loaded to overcome procrastination. Specifically, when $\beta b - c < 0 < b - c$, some incentive will be necessary to induce exercise at the daily level even once the startup costs are overcome. Together these results mean that for an agent facing simple one-time startup costs there is a tradeoff between the extensive-margin and intensive-margin responses to incentive structures.

The preceding paragraph highlights that even once people have crossed the extensive margin and joined the gym, without a sufficient additional incentive they might not exercise (i.e., when $\beta b - c < 0$). However, as Charness and Gneezy (2009) highlighted, habit formation may allow a temporary incentive to generate lasting changes in behavior.

Consider the naïve agent who has just joined the gym (e.g., has overcome any startup costs already).[7] Let h_t denote a level of "habit stock" for the activity at day t. The habit stock generates an additional utility gain ($h_t > 0$) or loss ($h_t < 0$) on days when the individual exercises. The net benefit of exercise from the day-t perspective is: $\beta b - c + h_t$. The stock of the habit is affected by the exercise decision on the prior day, rising when the person exercises and falling when she does not.[8] Assume, without loss of generality, that the initial habit stock on this first day after joining the gym is at a lower bound for habit stocks $h_0 = \underline{h} = 0$. Also, recall that we are considering the case of interest where the undiscounted net benefit of exercise is positive, but the discounted net benefit is negative ($\beta b - c < 0 < b - c$). If the habit stock can reach a critical threshold h^*, defined such that $\beta b - c + h^* > 0$, the individual will perceive positive net benefits of exercising each day and will exercise consistently.

An incentive may help the naïve agent establish this threshold habit. Specifically, she will choose to exercise on the first day iff:

$$\beta b - c + p_0 + \beta h_1' > 0, \tag{4}$$

where p_0 is an initial incentive. Here h_1' is the habit stock that will exist on day $t = 1$ if the agent exercises at $t = 0$ starting from the initial lower-bound of exercise. The term $\beta h_1'$ reflects the fact that the naïve agent perceives a benefit from today's exercise based on how it raises the habit stock (and hence utility for exercise she expects to do) for the following day.[9] A sufficient initial

[7] The logic in the preceding paragraphs on overcoming one-time startup costs does not change meaningfully if we consider this naïve agent recognizing the potential for habit formation at the same time she is trying to decide whether or not to join. As long as she believes she will begin exercising regularly starting tomorrow, then all that will matter for the extensive margin decision is whether she faces a sufficient spot incentive that day.

[8] We propose a specific formula for the habit-stock adjustment for use in simulations in Appendix A1.

[9] Note that the naïve agent believes she will begin exercising regularly starting the following day. So, she believes exercising today only brings forward this initial habit stock by one day.

incentive p_0 can induce the individual to exercise on the initial day. On the next day the incentive necessary to induce exercise will be (weakly) lower, since the increased habit stock will improve the net benefit of exercising for the individual. As this process continues, the habit stock eventually rises to h^*, and incentives will no longer be necessary.

If incentive programs could be targeted to the individual, the optimal approach would be to start the incentive just high enough to induce exercise and then reduce it in line with the growing habit stock. In a heterogeneous population with different parameters of utility and speeds of habit-stock adjustment, however, it is unclear ex-ante how to design a single optimal incentive structure. Different incentive structures will create tradeoffs. In particular, higher initial incentive payments can increase the share of people engaging in exercise. On the other hand, for a fixed incentive budget, the more front-loaded is the incentive structure the more rapidly incentives will need to fall over time. When habits build quickly, some degree of front loading may be beneficial to increase the share of people who initiate exercise, but if habits build slowly, a flatter profile of incentives may be more effective.

These results suggest that front-loading may be helpful for getting these naïve “procrastinators” to overcome one-time startup costs and make initial investments to establish an exercise threshold. Overcoming the one-time costs (i.e., extensive margin) depends on the degree of the initial spot incentive, while establishing a threshold habit will depend on whether incentives are elevated long enough to reach the threshold habit.

It is also important to consider how incentives will affect the behavior of non-procrastinators. In particular, there may be a set of people who have no long-run interest in using this gym. That is, consider a group for whom $b - c + \overline{h} < 0$, where $\overline{h}$ is the maximal habit stock that could be achieved through consistent exercise. This group does not want their future selves, let alone their potentially present-biased current selves, to exercise because even with habit formation, they are better off not exercising.[10] This group might, however, join the gym simply to receive incentive payments and hence we might denote them with the (hopefully not too pejorative) term “opportunists”.[11] While only the initial spot incentive mattered for the extensive-

[10] Note that in our context this may be a population who desire to exercise but not at the gym we study.

[11] Depending on the objective function of the principle (e.g., firm) offering the incentives, it may be valuable to use incentives to get this group to exercise even though it is not in their personal interest to exercise absent incentives.

margin decision of the "procrastinators", for these "opportunists" the total value of the available incentive payments can be crucial for determining whether they participate in the program on the extensive margin. For example, consider a case with no habit formation (for simplicity) and different incentive programs that offer the same daily spot incentive for different lengths of time. Denote the length of incentives in days as N. "Opportunists" will join the gym both if inequality (3) from above holds *and* if:

$$\beta b - c + p + (N-1)\beta(b - c + p) > f \qquad (5).$$

Inequality (5) will be more likely to hold for longer incentives (i.e., higher N). This leads to the interesting result that holding fixed the size of the daily incentive, longer programs will induce the same number of "procrastinators" to join but may induce more "opportunists" to join. These "opportunists", however, would stop exercising if the incentives were removed.

2.3 Habit formation in stochastic environments

Here we expand the simple framework to the case where habit stocks can grow and fade more smoothly by allowing for stochastic daily costs of effort (c_t). Recall our notation for the indicator of whether the individual exercised on a given day is $a_t \in \{0,1\}$. Assuming, for simplicity, that there are no startup costs (or they have already been paid), the individual will choose to exercise on a given day iff:

$$\beta b - c_t + h_t + p_t + \beta\big[E\big(V_{t+1}(h_{t+1}|a_t = 1)\big) - E\big(V_{t+1}(h_{t+1}|a_t = 0)\big)\big] > 0. \qquad (6)$$

The last term on the left-hand side of the inequality is the difference in the expected continuation value arising from today's exercise decision via its effects on tomorrow's habit stock level.[12]

The main insights about the tradeoffs between front-loaded and more constant incentives from Section 2.2 continue to hold in this expanded model as well, except incentives and habits now raise the *probability* of exercising on a particular day. If the maximal possible habit level ($\overline{h}$) is very large relative to the size of potential cost shocks, this stochastic framework can operate very similarly to the simpler threshold-habit-formation model discussed above where a sufficient

[12] The forward value function V_{t+1} is the value function for the time-consistent agent with $\beta = 1$. The naïve agent with $\beta < 1$, will wrongly believe this to be the value function reflecting her future decision process.

investment in exercise to reach $\overline{h}$ will lead to perpetual exercise. There are, however, at least two important additional insights we can generate from this richer framework when habituation levels are more modest – that is, when maximal possible habits are not large relative to the possible daily cost draws.

First, with modest habituation levels, in the absence of incentives, behavior will trend back to a baseline equilibrium even if incentives temporarily push the habit stock up to the full positive habituation level. Appendix Figure A2 in the modeling appendix (Appendix A.1) gives an illustration of these dynamics. We simulate the model (assuming a functional form for the distribution of daily costs and the habit-formation process) for the case of an 8-week incentive program similar to what we offer in our experiment. The key intuition is that peak habit stocks cannot be maintained indefinitely because they only raise the probability of exercising in Equation (6) but do not guarantee it. Poorer cost draws will cause the agent to sometimes not exercise and will hence erode the habit stock. As such, an incentive program can raise habit-stock levels and have some lasting effects on behavior for a while after it is removed, but behavior reverts to the pre-incentive baseline over time with the speed of this adjustment depending on how fast habits decay.

Second, in some cases it will be more effective to periodically raise and lower incentives over time (e.g., turning on and off) than to use a continual incentive program. Part of the intuition for this result comes from noting in Equation (6) that a raised habit stock (h_t) can substitute for an incentive (p_t). In addition, as habit stocks rise, at some point the impact of incentives on the probability of exercising will tend to fall.[13] This means that it can be effective to use an initial incentive to encourage exercise and then leverage the resulting increase in habit stock to lower the incentive in a subsequent period. The habit stock will fall without the incentive in place, creating a benefit to again raising the incentive. Appendix Figure A3 shows an example simulation of the model comparing a marginal incentive offered continuously for a period of time to the same marginal incentive instead offered every other week for twice as long. The troughs of exercise probability during the non-incentivized week for this periodic incentive are elevated relative to the baseline exercise probability and cumulatively are greater than the post-incentive elevation in

[13] The nature of this effect will depend on the distribution of cost shocks. With normally distributed cost, the CDF of the distribution is concave for probabilities above .5. If the level of $\beta b + h_t$ rises to the point where there is a high chance of exercising, additional incentives will generate a diminishing effect on the probability of exercise.

exercise seen with the continual incentive over that same period.[14] Whether this type of periodic incentive will generate higher overall average exercise depends on the parameters of the situation. For example, in Appendix Figure A4 we show an example where the average exercise generated by the periodic incentive is lower when the speed of habit formation (θ) is very low but higher when habits adjust at a moderate to fast pace. Deriving the optimal incentive scheme under modest habits in this stochastic environment is outside the scope of the present paper. However, we believe this exercise helps to highlight that thinking of habits as fluctuating stock of motivation rather than solely as a threshold level that guarantees behavior can open up important new considerations for incentive design, such as the use of fluctuating incentive levels.

3. Experimental Design

3.1 Setting

Our experiment took place at the headquarters of a Fortune 500 company in the Midwest. Roughly 2,000 employees work at this location, ranging from the call center to high-level management. The experiment operated in coordination with an onsite fitness facility, which is open from 6 a.m. to 6 p.m. Monday through Friday. The gym offers classes and standard aerobic and strength training machines. Gym membership is subsidized and costs employees $12.96 each 2-week pay period.

When employees enter the gym, they type their company identification number into one of two computer terminals inside the gym's main door entrance. These computerized login records provide high-quality data on visits made to the gym and provide both the data on which program incentives are based and our primary outcome measures. Importantly, participation in the study did not require employees to do anything different to log their visits. As is standard in many gyms, however, employees are not required to log out when they leave. As such, we do not have data on the length of time employees spend at the gym during a visit or on precisely what they do when they are there. We did not attempt to alter the check-in/out process at the gym because we wanted

[14] The effect in the non-incentivized weeks is due to two components. First the habit stock is raised from the prior incentivized period. Second, the agent is forward-looking and recognizes that there is a higher continuation value to the effect of today's exercise on habit stocks when there is an incentive in the future.

to keep the environment natural and worried that any new procedures might lead to differential compliance based on treatment status.

The reliance on check-in records may raise concerns that gym visits do not necessarily correspond to exercise since employees could earn incentives by visiting the gym without working out. We are not very concerned about this type of behavior for several reasons. First, there is a full-time staff at the gym who interact frequently with employees and monitor activity at the gym. This staff reported no changes in these types of activities or a rise in employees entering the gym without exercising. Second, Royer, Stehr, and Sydnor (2015) offered similar-sized incentives at this same site two years prior to this study and employed research assistants to unobtrusively monitor the gym on multiple days. The research assistants recorded no instances of employees checking into the gym without exercising. Finally, analysis of data from a follow-up survey, described below, indicates that a visit to the company gym was associated with around 45 minutes of exercise at the gym, too long to simply be a visit where the employee checks in and immediately leaves. A full description of this analysis appears in the Appendix, Section A.3.

3.2 Recruitment and Treatment Assignment

We recruited subjects from a list of all employees at the firm's headquarters by sending an email that contained a link to an online survey on wellness.[15] The email informed participants that they could earn a $25 gift card for completing both this initial survey and a short follow-up survey to be administered 8 weeks later. Gift cards were used throughout the study to simplify the distribution of incentives. We gave subjects a choice of gift cards to make them as fungible and close to cash as possible; the options included Amazon, Target, Wal-Mart and several local gas stations.

Respondents to our initial survey formed our subject pool for randomization, provided that they worked at the corporate headquarters and were physically able to exercise. We ran the experiment in 15 cohorts to ensure that the experiment did not overly tax the gym and gym staff. We enrolled the first cohort in June 2013 and the last cohort in April 2014. The closed nature of the corporate setting and the overlapping of cohorts raised the possibility for cross-talk between

[15] We exclude gym staff, executives, and human-resource staff who might have been aware of the details of this program. A copy of the survey is available upon request and will be included in an online appendix.

participants. We discuss this issue later and provide evidence that suggests communication between employees did not bias treatment effects.

Individuals were randomized into the treatments detailed in Table 1 and the randomization was stratified by whether they belonged to the company gym at the time of the initial survey. Both members and non-members were randomized into control, constant treatment, and kickstart treatment. We chose to incentivize up to 2 visits per week with only 1 visit per day eligible for incentives. Setting a cap higher than this would have shifted much of our incentive costs to inframarginal individuals (e.g., regular exercisers). Two visits per week however still allowed for incentivizing substantial changes in exercise given that among existing members, 65 percent attended 2 or fewer days per week prior to the study.

In addition to these two primary treatments, we offered two other targeted incentives to members and non-members separately. For the existing members this additional treatment, extended-sporadic, offered the same $10/visit incentive as the constant, but randomized the 8 weeks in which the incentive was received over a 16-week period. Each week, participants in this group received an email telling them whether or not the following week would be an incentivized week. The secondary treatment for the non-members offered them a $10/visit incentive but only over a 4-week period.[16] This constant-short incentive is the only one of the incentive schemes to have smaller total possible earnings – specifically half of the $160 maximum incentive earnings for the other programs. Finally, the non-members in all treatments received an additional $25 payment if they decided to join the gym during the incentive period, which was designed to offset the additional burdens they face in participating, including the need for a fitness assessment taking place at the gym as part of the gym-joining process.

This experiment took place in the same location as Royer, Stehr, and Sydnor (2015), which offered 4-week incentives to employees between February 2009 and March 2011. Roughly 78 percent of the employees in the company at the start of this experiment were employed at the company at the end of that earlier study, and although IRB regulations restrict us from matching

[16] We began the experiment assigning non gym-members to the extended sporadic treatment. After three cohorts, we received anecdotal reports that non-members were reluctant to join under this program because the incentives stretched over 16 weeks, requiring more gym-membership fees, while only incentivizing 8 of those weeks. Because of that difference in the nature of the effective treatments between the constant and extended sporadic incentive for the non-members, we discontinued allocating non-members to this treatment. We instead switched our secondary treatment for this group to be a 4-week incentive at the $10/visit rate.

the subject pools across experiments, some of the participants in the earlier study surely enrolled in this study. We think, however, that the effects of the earlier study on our findings here are likely minimal as the gap between the two experiments is two years. Additionally, the internal validity of our study is not compromised because randomization should lead to treatment assignment being orthogonal to previous participation.

The online follow-up survey collected self–reported measures of exercise type and duration to help us understand exercise patterns at and away from the company gym. Additionally, when picking up their earnings, individuals could earn $25 for a brief self-administered health assessment. Health measures included blood pressure, resting pulse, percent body fat and weight. We anticipated ex-ante that we would be underpowered to detect any differences in health measures based on treatment status and focus our analysis here on differences in visit patterns. We discuss the health measures in the Appendix, Section A.4.

Appendix Figure A1 summarizes the flow of subjects by treatment status. Of the 1,759 employees we contacted, 1,075 or 61% took the first survey and consented to participate in the study. Of those, 95% or 1,024 were randomized into treatment, with the others failing to meet inclusion criteria. Of the 1,024 subjects who were randomized to treatment, the final analysis sample includes 980 subjects with useable data.[17] Of those 980, 845 or 86% filled out the follow-up survey and 702 or 72% recorded their health measures. As we discuss in the Appendix, the lower response rate to the health measures creates scope for selection problems. The primary analysis of the paper, however, is focused on gym-attendance records, for which selection and attrition are not concerns.

Our final sample of subjects, 980, is fairly sizable. We based our sample size on power calculations for detecting in-treatment effects on gym attendance. Since ex-ante we expected that the differences between the control group and treatment groups would exceed those between the different treatment groups, we assigned more individuals to each of the treatment groups than to the control group. At 80 percent power, ex-ante, we had sufficient power to detect a 0.12 difference in the probability of visiting the company gym in a given week between any two treatment groups

[17] The number of subjects decreased to 1024 when we excluded those not medically able to exercise or who did not work at the site where the company gym is located. The number of subjects drops to 980 when we exclude those non gym-members in cohorts 1-3 who were assigned to the extended-sporadic treatment, which as noted in a preceding footnote was replaced with the constant-short treatment beginning with cohort 4.

and a 0.09 difference in this probability between the control group and any treatment group given baseline attendance rates and results from Royer, Stehr, and Sydnor (2015).

3.3 Descriptive Statistics

Table 2 presents descriptive statistics on the sample population, showing the overall mean and the mean for the control group for a number of characteristics measured with the initial survey. The table also provides a check that the randomization resulted in balance by providing the difference in these values for each treatment group relative to control and the p-value for a joint test that the mean of each variable is the same for all groups. The table is split into two panels to separate participants by their ex-ante company gym membership status, which we stratify our randomization on. Overall, the randomization is balanced on observable characteristics. Only one of 26 p-values testing whether the pre-treatment means are equivalent across groups is at or below 0.05.

Both the samples of members and non-members are roughly equally split along gender lines. The sample is well-educated; over 70 percent have college degrees. Rates of overweight and obesity are high, although our overall rate of 62 percent is a bit below the overall national average of 70.7 percent. The slightly lower rate in our sample is not surprising given the negative correlation between education and rates of overweight/obesity. Interestingly, despite different rates of exercise, members and non-members have similar rates of overweight/obesity.

In the initial recruitment survey, we inquired about individual exercise patterns – overall, at the company gym, at any other gym, and outside of the gym. Such data are useful for assessing how well a company gym based intervention might work for changing exercise behaviors. For members, most of their exercise takes place at the company gym (average of 2.3 days/week out of 3.57 days/week). Non-members exercise less: 2.24 days/week, with 74% of this exercise occurring outside of a gym and 26% occurring at other gyms, outside the workplace. Non-members and members alike appear to have a desire to increase their exercise as their target level of exercise exceeds their actual exercise, with non-members reporting a lower subjective probability of reaching their self-chosen targets.

4. Results

4.1 Participation Decisions of Non-Members

We divide our main analysis in two sections – first considering the effects among employees who were not members of the company gym prior to the experiment (i.e., non-members) and then the effects among members. For non-members, startup costs, as predicted from the model, are likely to be critical in their decision of whether or not to join the gym when an incentive is introduced.

We estimate the effects on the extensive margin for non-members using the following OLS regression model that includes treatment-group dummies and controls for observable characteristics:

$$Joined_i = \alpha + \delta_1 D_i^{constant} + \delta_2 D_i^{kickstart} + \delta_3 D_i^{constant-short} + X_i\beta + \sum_{j=1}^{14} \mu_j I_i^j + \epsilon_{it}, \qquad (7)$$

where $Joined_i$ is a binary indicator equal to 1 if an employee attends the gym at least once during weeks 1-8 where week 1 is defined as the first incentive week, and D_i^x are the 3 treatment indicators for the 8-week constant incentive program, the 8-week kickstart program, and the 4-week constant-short program. I_i^j are binary indicators for the 15 cohorts over which subjects were recruited and randomized. The controls included in X_i are male, age, married, a set of dummies for educational attainment, and a set of dummies for level of interest in joining the company gym. All of these variables were collected in the initial survey, prior to treatment assignment. Additional controls are not strictly necessary in a randomized trial, and we find they have small effects on the estimated treatment effects, but they do help to increase the precision of the estimates.

Table 3 displays the regression results. Before discussing in detail these results, it is important to note that in the absence of incentives, 4.6% of the control group joined the gym over weeks 1 through 8. This is certainly not surprising as some will naturally join over time and moreover, does not imply that the incentives had effects on non-incentivized individuals. In other non-reported analyses, we find that this joining rate is roughly similar to the join rate of non-members during the pre-experiment period.

Looking at the treatment effects, both the constant and kickstart treatments had large effects on the probability of joining, increasing the probability by 10 percentage points relative to the control group. In levels, this implies that the join rate was approximately 3 times higher for these two treatment groups relative to control. The constant-short incentive works less well; the point estimate is 0.032 (roughly one-third the size of the other two incentive groups) and is statistically insignificant. The p-value of a Wald test for the equality of the effects of the constant and constant-short incentives is 0.06.

Contrary to the prediction of the model for "procrastinators", we find no evidence that the front-loaded kickstart treatment increased join rates more than the constant treatment. Within the context of our model, there could be two explanations for this lack of a differential response. First, there simply may not be very many procrastinators in the population. Second, among the procrastinators there may be diminishing returns to spot incentives, at least in the range of our incentive size. That is, the distributions of present-bias and fixed costs in the population may result in a share of procrastinators who can be motivated by a $10 incentive, but few additional procrastinators who would respond only for higher incentives between $10 and $25. This does not rule out that there may be a share of procrastinators who would respond to a much higher front-loaded spot incentive above $25.

The finding that the constant-short incentive induced fewer people to join than the longer constant incentive is consistent with model predictions for "opportunists" but not for "procrastinators". That is, the model predicts that the "procrastinators" respond to the peak spot incentive and should respond similarly to the two constant programs. The marginal joiners under constant compared to constant-short should be the "opportunists" who do not intend to begin a long-term habit of using the gym and are responding mostly to the total value of the incentive program. We see some suggestive evidence consistent with that interpretation when we look at the join rates for non-members split by how interested they said they were in joining the gym in our pre-survey. Among the 19% of non-members who stated they were probably or definitely interested in joining the gym, the join rate among constant-short was 26%, compared to 24% for kickstart, 33% for constant, and 14% for control. While these are small-sample cuts, and this analysis was not pre-specified, they suggest that the constant-short scheme was fairly comparable to the longer incentives in terms of motivating this group. Among the 81% who were somewhere between uncertain and definitely not interested in joining, the join rate for constant short was 6% compared to 13% for kickstart, 10% for constant and 2% for control. This is at least suggestive that most of the additional join rate for the longer incentives came from those with low ex-ante interest in using this gym.

Another way of exploring motivation on the extensive margin is to consider how the different incentives affected join rates based on individuals' self-reported exercise frequency prior to our interventions. In columns (2) and (3) of Table 3, we show the estimates of Equation (7) for two subsamples of non-members: 1) those who reported exercising two or fewer days in the week prior to the initial survey and reported that they typically exercise two or fewer days per week ("low exercisers"), and 2) those who reported three or more days in the week prior to the survey and reported that they typically exercise three or more days per week ("high exercisers").[18] The results show that all the incentives had similar effects among ex-ante "high exercisers", inducing 11 to 13% more of these people to join than in control. The difference in join rates between the constant-short and longer incentives was concentrated among the "low exercisers". The constant-short induced no more people to join from this group than control and we can clearly reject the hypothesis that constant and constant-short incentives had the same effect (p=0.01). This split was not pre-specified, so it is important to have some caution in interpreting these results. However, we believe this cut of the data is natural for heterogeneity considerations, especially since companies may be interested in how different incentive programs motivate participation for those who are both not using the company gym and not exercising much elsewhere.

Overall, these results create an interesting tension when considering the value of longer incentive programs. On the one hand, the marginal joiners may be "opportunist" types who have little perceived personal value for additional exercise. On the other hand, many of these people may be those who are physically inactive to begin with and the incentive may be more meaningfully changing their exercise behavior. We discuss this issue more in our conclusion

[18] Our categorization of low and high exercisers takes advantage of two different measures of self-reported exercise collected in the initial survey, one about days of exercise last week, and one about exercise frequencies in a typical week. For both measures, the median value reported is 2. Our findings are very similar if we instead use either measure alone to categorize exercise frequency.

section. We also return to this issue below when we analyze the post-incentive effects of the treatments.

4.2 Intensive Margin Effects for Non-Members

In this section we analyze the visit patterns over time for the non-members in our study. Figure 1 provides a first look at the patterns of behavior showing a LOESS-estimation smoothed plot of the fraction of non-members making at least one visit each week to the gym. The figure stacks different cohorts and shows data from 9 weeks prior to our intervention, when none of these people used the gym, to 16 weeks after the start of the incentive period. The figure confirms the results discussed in the prior section on the extensive-margin responses, showing that the constant and kickstart treatments had similar effects, both peaking in week 2 with around 14% of non-members making a visit to the gym. The constant-short group had lower peak visits. A few new notable patterns emerge, though, from this visual analysis. First, we see that the fraction using the gym diverges between constant and kickstart in the second half of the incentive period when the incentives for kickstart were only half as high ($5 vs $10). Second, all groups see sharp declines in use after the end of the incentive period, but the fraction of people who continue using the gym after incentives is modestly higher than in the control. Importantly, these post-incentive effects appear similar across the incentives.

We formally estimate these effects using regression analysis. Throughout, we consider two measures of visit rates: *Any Visit*, a binary variable indicating at least one visit in a given week (consistent with Figure 1), and *Number of Visits*, the number of days with a visit in a given week. We organize the analysis into three periods – weeks 1-4, 5-8, and 9-16. As the constant-short treatment lasts just four weeks, we break the first eight weeks into two four-week periods. We again include the controls we used earlier for the extensive margin responses for non-members. Using data for weeks 1-16, our regressions to estimate the treatment effects take the following form:

$$y_{it} = \alpha + \sum_{k=1}^{3} P_{i,t}^{k}\left(\gamma_k + \delta_{1,k} D_i^{constant} + \delta_{2,k} D_i^{kickstart} + \delta_{3,k} D_i^{constant-short}\right) + X_i\beta + \sum_{j=1}^{14} \mu_j I_i^j + \varepsilon_{it} \quad (8)$$

where y_{it} is the person-by-week visit measure, D_i^x are the 3 treatment indicators (constant, kickstart, and constant short), and $P_{i,t}^k$ are three time-period indicators (weeks 1-4, weeks 5-8, and weeks 9-16, respectively). The rest of the variables are defined as before in equation (7). The parameters $\delta_{1,k}, \delta_{2,k}, \delta_{3,k}$, our coefficients of interest, estimate the treatment differences for each of the treatment groups (constant, kickstart, and constant-short respectively) relative to control in each of the time periods.

We display the regression results in Table 4. Column (1) provides the results with any visit as the dependent variable whereas column (2) shows the results with the number of visits as the dependent variable. The column (1) estimates confirm the visual patterns seen in Figure 1. In weeks 1-4, kickstart and constant work equally well – increasing any visits margin by 0.11 and the number of visits by 0.25-0.28. After that period, the effects begin to diverge slightly. From Figure 1, it is evident that kickstart has a declining effect over time; the constant group's patterns are flatter. The constant effect in weeks 5-8 is 0.10 whereas for the kickstart it is 0.07 for any visits. The effects during the post-incentive period using the any visits measure for these two are roughly equal in magnitude; 0.036 for constant and 0.030 for kickstart. In line with the differences found in the participation decision, the treatment effects for constant are nearly twice as large as those for constant-short in weeks 1-4, although the difference is not statistically significant. The initial post-treatment effects for constant-short (weeks 5-8) are 0.027, falling to 0.017 in weeks 9-16. These effects are broadly similar to, and statistically indistinguishable, from the persistence effects for the longer constant and kickstart treatments. The results in column (2) for number of visits show the same qualitative patterns.

The relative similarity of the persistence of effects across treatments despite differences in join rates and visit patterns during the incentive period speaks to the predictions of the model. In particular, these patterns are consistent with model predictions for the case with startup costs that the marginal joiners for the longer programs are "opportunist" types with fairly little interest in using the gym. It does not appear that the longer programs induced substantially stronger habits for this group that overcame their lack of interest in using this gym.

4.3 Intensive Margin Effects for Members

We now compare the effects of the three incentive structures given to existing gym members. For this group, all three incentive designs had the same budget but spread the possible earnings out in three different ways, as described in Table 1. Similar to Figure 1, Figure 2 stacks the cohorts and shows the LOESS-estimation-smoothed time trend of the fraction of this group using the gym at least once in each week. The graph includes a 6-week pre-period beginning a month prior to our interventions and the 16 weeks starting with the first incentivized week; we exclude the month around when we were recruiting subjects and informing them about the incentives. A few notable visual patterns emerge. First, all incentives meaningfully increase the fraction using the gym. Second, the peak effect for the kickstart in the early weeks is quite similar to the effect for the constant incentive. Third, the kickstart treatment's visit rates fall from that peak significantly over the period where the incentives were lowered to $5/visit. Fourth, the constant incentive shows elevated visit rates after week 8 when incentives ended relative to control, but the kickstart does not. Finally, the extended-sporadic has more level visits rates over 16 weeks and these rates appear to be more than half of the incentive-period effect of the constant treatment.

We quantify these patterns following the same regression setup as equation (8) with four alterations: 1) we replace the constant-short treatment with the extended-sporadic treatment, 2) we organize the data into two periods - weeks 1-8 when all incentive treatments were active and weeks 9-16 when only the extended-sporadic group was still eligible for incentives, 3) we include individual fixed effects, instead of *X*s and cohort dummies since we have pre-incentive company gym attendance and thus parameters γ_k estimate the visits rate differences for the control group in these three different time periods relative to their pre-treatment visit rates and our main treatment effects ($\delta_{1,k}, \delta_{2,k}, \delta_{3,k},$) are essentially difference-in-difference estimates, and 4) we estimate our regressions using data from weeks -9 to -4 and 1-16, excluding weeks -3 to 0 when subjects were being recruited for the intervention.[19]

Table 5 displays these regression estimates for the outcomes *Any Visit* and *Number of Visits*. Focusing on columns (1) and (3), the regression estimates largely confirm the visual pattern from Figure 2. Over the 8-week incentive period, the constant incentive elevated the fraction visiting relative to control by 18 percentage points from a base of 62 percent (a 29 percent

[19] As expected, the regression estimates are quite similar if we do not replace the *X*s and cohort fixed effects with the individual fixed effects. The individual fixed effects, however, are useful as a means of variance reduction.

increase), while the kickstart incentive increased it by a more modest 12 percentage points. If we estimate the treatment effects over weeks 3-8 (not reported in Table 5), when the constant incentive was $10/visit and kickstart was $5 for visits, we observe an 18 percentage-point effect for constant incentive and 9 percentage-point effect for kickstart, with a p-value of 0.03 on the difference in those effects. In contrast, the treatment effect in weeks 1 and 2, when the kickstart incentive was $25, were only modestly higher for the kickstart (21 percentage points vs 18 percentage points) and not statistically significantly different. These results suggest strongly diminishing returns to the spot incentives among the existing members.

The results from weeks 9-16 provide further evidence of the constant incentive's greater efficacy. This group showed a statistically significant 9 percentage point increase in *Any Visit* in the first eight weeks after the incentive period, an effect half the size of the treatment effect during the incentive period. The kickstart group, by contrast, exhibits no effect in this post-period. Overall, taking into account both the in-treatment and post-treatment effects, the constant incentive scheme outperforms the kickstart incentive. The average difference in the treatment effects on any visits between constant and kickstart across weeks 1-16 is 0.075 (p-value 0.025). However, the differences are more muted and not statistically significant using the number-of-visits measure (column 3), though we note also that the number of visits is a noisier measure.

The patterns in Figure 2, however, also suggest that the persistence effects for the constant treatment are declining over time. Our regression estimates confirm that pattern. For example, the estimate over weeks 17-24 (not shown in Table 5) for *Any Visit* is only 0.055 (compared to 0.093 for weeks 9-16) and the estimate is essentially zero for the *Number of Visits* measure (-0.02).

Comparing the results of the extended-sporadic incentive to the constant incentive in columns 1 and 3 show patterns consistent with this periodic incentive leveraging temporary habit formation. In the absence of habit-formation effects, we would expect that the extended-sporadic would show half of the treatment effect of the constant treatment during weeks 1-8, since it incentivized half as many weeks. However, the model with habit formation predicts that the average effect will be higher because visit rates will be boosted relative to control also in the non-incentivized weeks for the extended-sporadic treatment. We estimate effect sizes for extended sporadic in weeks 1-8 that are 60 percent of those for constant using the *Any Visit* measure and 100 percent as effective using the *Number of Visits* measure. In weeks 9-16, when the constant and kickstart groups no longer earn incentives but the extended sporadic group continues to be incentivized, the extended sporadic group's treatment effect appears even higher than during weeks 1-8. This indicates that the effect of the weekly incentives is not diminishing over time, and on the contrary, may be increasing, consistent with the notion of habit formation. The overall average effects on visit rates across weeks 1-24, a 6-month period that captures the incentive periods and at least 8 weeks post-incentives for both treatments, however, are similar between constant and sporadic, with almost identical effects on *Any Visits* and slightly higher (though statistically insignificantly different) rates for extended-sporadic using the *Number of Visits* measure.

We can further explore the habit formation process by contrasting behavior in incentivized versus non-incentivized weeks of the extended sporadic treatment. Columns (2) and (4) of Table 5 add to the regression model interactions between the sporadic incentive indicator and dummy variables for whether or not week *t* was incentivized for member *i*. According the model, in non-incentivized weeks, current habit stock acts as a substitute for incentives to motivate continued attendance. Thus, during those unincentivized weeks, one might expect increased attendance relative to the control group but that attendance should be less than during the incentivized weeks. This bears out in the data. During the non-incentivized weeks for the extended sporadic, there is a positive treatment effect of 0.065 to 0.11 on *Any Visit*, depending on the time period, and for weeks 9-16, this effect is statistically significant. The incentivized-week effects for the extended sporadic group during both weeks 1-8 and 9-16 are roughly as strong as those for the constant group during its incentive period of weeks 1-8. For the number of visits measure, the extended sporadic effects actually exceed the constant effects.[20]

Overall, these results are consistent with the model predictions regarding habits that do not reach a self-sustaining threshold. In particular, the extended-sporadic treatment shows promising results consistent with the model's predictions. However, this incentive program may leverage additional mechanisms. For example, the sporadic nature of the incentive may have made it more salient. The extended-sporadic design also involved contacting participants more frequently, via weekly notifications of the following week's incentive status. This additional contact might have

[20] This dampens any concern that the on-and-off nature of the incentives confused subjects, and if anything, suggests instead that sporadic incentives maybe more salient than consistent incentives.

served as reminders, which could increase the effectiveness of this treatment and may explain why attendance was elevated in non-incentivized weeks relative to the control. However, the reminder effect likely explains only a very small amount of the extended sporadic treatment effect. Calzolari and Nardotto (2017) document that weekly reminders sent to gym members via email resulted in an increase in visit rates of 3 percent in their study. A reminder effect of this size for our members would predict an increase of 0.06 visits per week for our sample. This is small relative to both the overall effect of sporadic (0.43-0.49 visits per week) and the effects observed during the non-incentivized weeks for this treatment (0.31-0.38 visits per week). Reminder effects, could though, plausibly explain a sizeable share (about half) of the increase in effect we observe for the incentivized weeks of the extended-sporadic treatment relative to the incentive period for the constant incentive looking at the *Number of Visits Measure*.

In Table 6 we provide estimates of the heterogeneity of the effects among members by dividing members into two groups: 1) those who exercised 2 or more times per week for at least half of the weeks from week -9 to week –4, and 2) those who exercised 2 or more times per week for less than half of the weeks from weeks -9 to -4. In this sense, the "high-exercise" group consists mainly of those who are already exercising 2 or more days a week and for whom the incentives should have little impact on their behavior. Consistent with the fact that our incentives were capped at 2 visits/week and hence were inframarginal for high-exercisers, the ex-ante low-use members are responsible for the significant treatment effects in Table 5. For the low-use members, the effects of all incentive schemes for weeks 1-8 are large and statistically significant at the 5 percent level. In contrast, fewer of the effects (4 out of 6) are statistically significant at the 5 percent level for high-use members. For the high-use members, the concern that the treatment effects might be negative if financial incentives crowd out intrinsic motivation turns out to be unjustified.[21] Consistent with the findings of Charness and Gneezy (2009) and Sen et al. (2016), financial incentives do not appear to harm intrinsic motivation in this setting or to otherwise create a negative target effect for those who are already heavily engaged in exercise.

[21] The only coefficients that appear significantly negative are those for number of visits for all time periods. This simply reflects mean reversion (recall that we have defined this subgroup as employees with above-median visits in the pre-period, which is the omitted period from the regression). Mean reversion can also explain why the trend in the control group's gym visits was downward for members as a whole, since some current members might have quit their memberships during our study period but remained classified as members within our study.

4.4 Cost Efficiency of the Treatments

In addition to knowledge about the treatment effects, policy makers may also be interested in cost effectiveness. We calculate cost effectiveness by dividing the treatment effect (measured over weeks 1-24 of the experiment) by per-subject cost of the incentive treatment and also by expressing the effect in terms of dollars spent per additional exercise visit induced. We present these estimates in Table 7.

Among members, the extended sporadic is the most cost effective followed by the kickstart and then the constant.[22] The incentive programs are more cost-effective among the non-members because they make far fewer infra-marginal visits (i.e., most of the visit payments go to visits that would not have happened in absence of the incentive program). The results point to an important feature of incentive design: being able to target incentives to marginal individuals (i.e., those whose behavior is altered by the incentive scheme) has an important impact on the cost-effectiveness of the incentive program.

5. Robustness checks

In the Appendix we conduct a series of robustness checks and additional analysis that help support the interpretations of our experimental results. Most of the issues we address are unrelated to our primary focus of comparing how different incentive designs affect behavior, but they are relevant for how we interpret what observed visit patterns mean in terms of real exercise behavior. For example, we consider the extent to which people might be substituting from other exercise when using the company gym. This issue is likely most important for non-members prior to the study. Based on self-reports of behavior, we estimate for ex-ante non-members, approximately 79% of the gym days induced by the incentives represent new exercise. A second issue is whether people are actually exercising when they visit the gym. We analyze survey data on exercise activities and conclude based on that evidence that people are likely exercising in similar ways when they visit the gym under incentives as when not under incentives, typically involving workouts of around 45 minutes.

[22] The kickstart appears more cost effective than constant in this table largely because that group showed a statistically insignificant rise in number of visits far after the treatment period ended, which we attribute mostly to noise.

Another issue that could meaningfully affect the interpretation of our results is whether awareness of the treatment patterns, or generally the effects of being in a study, are affecting behavior of our control group or spilling over across treatment groups. This is a concern in most field experiments and especially in closed environments like a workplace. However, a number of pieces of evidence suggest that this type of cross-contamination was minimal.

First, since cross-contamination is most likely to occur between employees who work together, we can examine the overlap within cohorts of employees in the same organizational unit, noting the company has 268 different units. Across all cohorts, 73 percent of subjects were the lone employee from their unit in their cohort. Second, the treatment effects do not vary with the number of employees from one's unit in the experiment. If the incentives were more effective when one's co-worker was also in the experiment, we would expect that the treatment effects would have positive interactions with the number of co-workers in the experiment.

Third, the introduction of incentives does not appear to affect the behavior of the control group. We assess this in two ways: 1) by examining the change in usage/membership for control group subjects during the intervention relative to before the intervention and 2) by comparing the patterns of visits/membership of non-members before the intervention when no incentives were in place with the analogous patterns of control group non-members during the intervention. Starting with the first approach, if we pool members and non-members, we observe no significant trends in gym use in the control group from the pre-treatment period to the main treatment periods.

Following the second approach, we compare the joining rate of non-members in the control group during weeks 1-8 of the experiment to a joining rate prior to the intervention. To do this we isolate subjects who were not gym users over the period -24 to -19 weeks. We then measure the fraction who attended the gym at some point between weeks -14 to -7. These periods occur prior to our contact with the subjects about the study and thus cannot have been influenced by it. We focus on these periods because they are comparable to that used in measuring joining the gym in the experiment (i.e., the gap between the pre-period and the joining period is the same as in our actual measure (5 weeks) and the length of the joining period is the same (8 weeks)). The join rate for the pre-period was 3.5% which is similar to the 4.6% join rate for the non-members in the control group. We also see similar *Any Visit* and *Number of Visit* rates during our treatment period (weeks 1-8) for the non-members in our control group relative to this earlier group of joiners.

Overall, these pieces of evidence suggest that any bias in our estimates for non-members from cross contamination is likely to be small.

6. Discussion & Conclusion

This study tested the effects of differently structured incentives to exercise using a large field experiment in a workplace setting. We found that relative to a constant incentive, using variable incentives that front-loaded incentive payments did not improve overall participation in the program and was less effective at sustaining exercise over time. However, periodic incentives, where incentives are turned on and off over time, showed more promising effects. While the exact treatment effects we estimate here are specific to the population and environment we study, we think these results have some broader implications.

First, the experimental results are not encouraging for the use of front-loaded incentives. We observed no additional extensive-margin response to a $25 incentive versus a $10 incentive. Of course, this does not rule out that a much higher front-loaded incentive could have had a stronger effect. However, we also see that smaller incentives following the front-loaded period have weaker effects than the $10 incentive, so there will likely be a strong tradeoff to trying more sharply front-loaded designs. It is also possible that front-loading would be more effective in other contexts or for other behaviors. Our theoretical model predicts that front-loading is more valuable in settings where a large fraction of otherwise-motivated people are procrastinating on initiating a behavior for which self-sustaining habit thresholds can be established quickly. In such settings, it may be useful to experiment with different versions of front-loading to optimize treatment responses.

Second, our findings show that a longer and more lucrative incentive program can increase participation on the extensive margin, but the marginal participants might be quite different from those already willing to participate. The theoretical model suggests that most of these additional joiners would be “opportunist” types who perceive low value to the behavior and do not intend to continue afterwards. Our experiment finds some evidence consistent with that prediction. On the other hand, the fact that the marginal joiners had lower exercise rates at baseline suggests that it is important to consider the objective function of the principal. If the objective is to get low exercisers more physically active, perhaps due to paternalistic concerns or the externalities that low physical activity might have via health care costs, then a larger incentive program that broadens participation might be valuable. If instead the goal is to help people engage in an activity they want to do for themselves, or to generate lasting change in their behavior, there may be a limited benefit to extending the length or total earnings of the program.

Third, our study provides important new evidence on the limited ability of temporary incentives to generate persistent new habits. This experiment tested a long incentive period (8 weeks) relative to many of the existing studies. In particular, the constant treatment here is similar to a 4-week $10/visit incentive tested at the same worksite a few years earlier by Royer et al. (2015), with the primary difference being that the prior study paid for 3 visits per week instead of 2. The Royer et al. (2015) study found small and statistically insignificant post-incentive effects among existing members of the gym, even though a large fraction of members reported exercising less than they desired. Measuring persistence as the ratio of the 4-week post-incentive effect to the 4-week incentive effect, they found persistence of 21%, which was significantly lower than the 50% persistence that Charness and Gneezy (2009) documented in an experiment with college students. One possibility highlighted by Royer et al., was that the incentive period was too short to induce strong habit formation. Consistent with that point, the 8-week incentive here generated initial persistence of 50%, in line with the original Charness and Gneezy findings. However, unlike Charness and Gneezy, in this setting we see that this initial persistence fades sharply over time. While our results do not rule out that an even longer incentive could have generated more persistent effects, they call for some caution about the viability of using a temporary incentive to generate lasting changes in exercise, especially for working-age adults.

Fourth, both the theoretical model and empirical results in our study suggest that there may be value in exploring the use of periodic incentives. The benefits of periodic incentives will likely be higher in situations where people face shocks that erode positive habits over time. In our setting of exercise among working adults, interruptions due to injuries, travel, work, or family constraints are common. In other settings, like establishing a hand-washing habit (Hussam et al., 2017), habit erosion may be less of a concern. Our sporadic-incentive treatment provides a novel look at the potential benefits of using periodic incentives that leverage temporary habit formation. We hope future research will explore this issue in more depth.

Finally, in considering how incentives interact with habit formation and startup costs, we focused on a relatively simple model of naïve present bias. Other behavioral forces, however, may coincide with time inconsistency, e.g. partial sophistication about one's present bias, projection bias from current states to future states, and fundamental biases and mispredictions about habit-formation processes. All of these issues may affect how incentives work for behaviors like exercise. We suspect that a promising direction for continued research in this area will be to gather more data about the beliefs people have about how their habits will develop over time.

References

Acland, Dan and Matthew Levy. 2015. "Naiveté, Projection Bias, and Habit Formation in Gym Attendance." *Management Science,* 61(1): 146-160.

Babcock, Philip and John Hartman. 2010. "Networks and Workouts: Treatment Size and Status Specific Peer Effects in a Randomized Field Experiment." NBER Working Paper 16581.

Babcock, Philip, Kelly Bedard, Gary Charness, John Hartman, and Heather Royer. 2015. "Letting Down the Team? Evidence of Social Effects of Team Incentives." *Journal of the European Economic Association,* 13(5): 841-870.

Bachireddy, Chethan, Leslie John, Katherine Milkman, Francesca Gino, and Bradford Tuckfield. 2017. "How Can We Optimally Reward Exercise and Build Lasting Habits? A Field Experiment." Working paper.

Becker, Gary and Kevin Murphy. 1988. "A Theory of Rational Addiction." *Journal of Political Economy* 96(4): 675-700.

Beshears, John, James J. Choi, David Laibson, and Brigitte C. Madrian. 2017. "Does Front-Loading Taxation Increase Savings? Evidence from Roth 401 (k) Introductions." *Journal of Public Economics* 151: 84-95.

Burman, Leonard E., William G. Gale, and David Weiner. 2001. "The Taxation of Retirement Saving: Choosing between Front-loaded and Back-loaded Options." *National Tax Journal,* Sep 1: 689-702.

Calzolari, Giacomo and Mattia Nardotto. 2017. "Effective Reminders." *Management Science* 63(9): 2915-2932.

Carrera, Mariana, Heather Royer, Mark Stehr and Justin Sydnor. 2018. "Can Financial Incentives Help People Trying to Establish New Habits? Experimental Evidence with New Gym Members" *Journal of Health Economics,* 58: 202-214.

Cawley, John and Joshua Price. 2013. "A Case Study of a Workplace Wellness Program That Offers Financial Incentives for Weight Loss." *Journal of Health Economics,* 32(5): 794-803.

Charness, Gary, and Uri Gneezy. 2009. "Incentives to Exercise." *Econometrica,* 77(3): 909–931.

Daly, Michael, Colm P. Harmon, and Laim Delaney. 2009. "Psychological and Biological Foundations of Time Preference." *Journal of the European Economic Association.* 7(2-3): 659-669.

Finkelstein, Eric, Laura Linnan, Deborah Tate, and Ben Birken. 2007. "A Pilot Study Testing the Effect of Different Levels of Financial Incentives on Weight Loss among Overweight Employees." *Journal of Occupational and Environmental Medicine* 49(9): 981-989.

Gneezy, Uri, Stephen Meier, and Pedro Rey-Biel. 2011. “When and Why Incentives (Don’t) Work to Modify Behavior.” *The Journal of Economic Perspectives* 25(4): 191–209.

Haisley E, Volpp KG, Pellathy T, Loewenstein G. 2012. “The Impact of Alternative Incentive Schemes on Completion of Health Risk Assessments.” *American Journal of Health Promotion,* 26(3): 184–188.

Halpern, SD, B French, DS Small, K Saulsgiver, MO Harhay, J Audrain-McGovern, G Loewenstein, TA Brennan, DA Asch, KG Volpp. 2015. “Randomized Trial of Four Financial-Incentive Programs for Smoking Cessation.” *New England Journal of Medicine,* 372(22): 2108-2117.

Hussam, Reshmaan, Atonu Rabbani, Giovanni Reggiani, and Natalia Rigol. 2017. “Habit Formation and Rational Addiction in Handwashing.” Harvard Business School Working Paper 18-030.

John, Leslie K, Loewenstein, George, Troxel, Andrea B, Norton, Laurie, Fassbender, Jennifer E., and Kevin G. Volpp. 2011. “Financial Incentives for Extended Weight Loss: A Randomized, Controlled Trial.” *Journal of General Internal Medicine,* 26(6): 621-626.

Just, David. R., and Joseph Price. 2013. “Using Incentives to Encourage Healthy Eating in Children.” *Journal of Human Resources* 48(4):855-872.

Karlan, Dean., Margaret McConnell, Sendhil Mullainathan, and Jonathan Zinman 2016. “Getting to the Top of Mind: How Reminders Increase Saving.” *Management Science, 62*(12), 3393-3411.

Loewenstein, George, Joseph Price and Kevin G.M. Volpp. 2016. “Habit Formation in Children: Evidence from Incentives for Healthy Eating.” *Journal of Health Economics,* 45: 47-54.

Milkman, Katherine L, Julia A Minson, and Kevin G.M. Volpp. 2014. “Holding the Hunger Games Hostage at the Gym: An Evaluation of Temptation Bundling.” *Management Science* 60(2): 283-299.

Mochon, Daniel, Janet Schwartz, Josiase Maroba, Deepak Patel, and Dan Ariely. 2017. “Gain Without Pain: The Extended Effects of a Behavioral Health Intervention.” *Management Science,* 63(1): 58-72.

O’Donoghue, Ted, and Matthew Rabin. 1999. “Doing It Now or Later.” *American Economic Review,* 89(1): 103-124.

Patel MS, Asch DA, Rosin R, et al. 2016. “Framing financial incentives to increase physical activity among overweight and obese adults: a randomized, controlled trial.” *Annals of Internal Medicine,* 164(6): 385–394.

Royer, Heather, Mark Stehr, and Justin Sydnor. 2015. “Incentives, Commitments, and Habit Formation in Exercise: Evidence from a Field Experiment with Workers at a Fortune-500 Company.” *American Economic Journal: Applied Economics*, 7(3): 51-84.

Sen, Aditi, David Huffman, George Loewenstein, David A. Asch, Jeffrey T. Kullgren, Kevin Volpp, “Do Financial Incentives Reduce Intrinsic Motivation for Weight Loss?: Evidence from Two Tests of Crowding Out” (2016) *Nudging Health: Health Law and Behavioral Economics*, edited by I. Glenn Cohen, Holly Fernandez Lynch, and Christopher T. Robertson.

Tagney, June P., Roy F. Baumeister, and Angie Luzio Boone. 2004. “High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success.” *Journal of Personality*, 72: 271-324.

Thaler, Richard H. 1994. "Psychology and Savings Policies." *The American Economic Review* 84(2): 186-192.

Volpp, Kevin G., Leslie John, Andrea Troxel, Laurie Norton, Jennifer Fassbender, and George Loewenstein. 2008. “Financial Incentive-Based Approaches for Weight Loss: A Randomized Trial.” *Journal of American Medical Association* 300(22): 2631-2637.

Volpp, Kevin G., Andrea B. Troxel, Mark V. Pauly, Henry A. Glick, Andrea Puig, David A. Asch, Robert Galvin et al. 2009. "A randomized, controlled trial of financial incentives for smoking cessation." *New England Journal of Medicine* 360(7): 699-709.

Figure 1. Fraction of Non-Members with Gym Visits by Treatment Group and Week

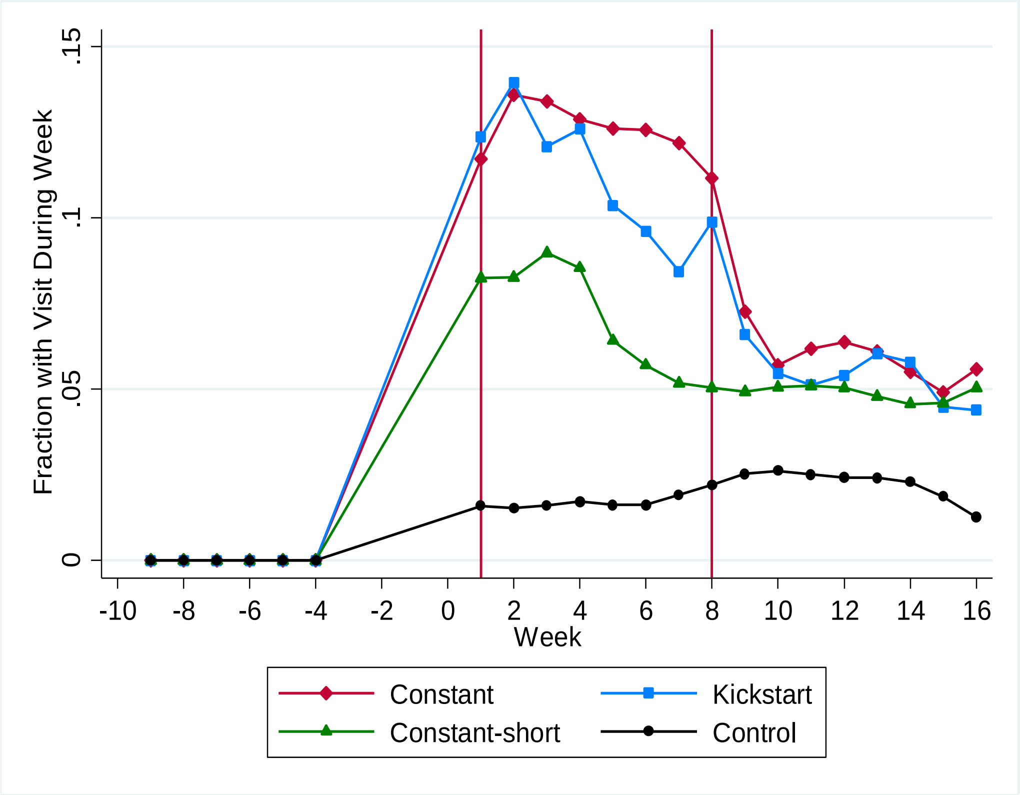

Figure notes: Figure presents LOESS smoothed plots with bandwidth set at 0.4. Outcome variable is an indicator for making at least one visit to the company gym during that week. Week 1 (marked with vertical line) denotes the first week of the incentive program. Pre-treatment visits rates for weeks -9 through -4 are included in the graph. Weeks -3 through 0 are omitted because these were weeks in which the initial survey and/or information about the treatment assignment was sent to subjects.

Figure 2. Fraction of Existing Members with Gym Visits by Treatment Group and Week

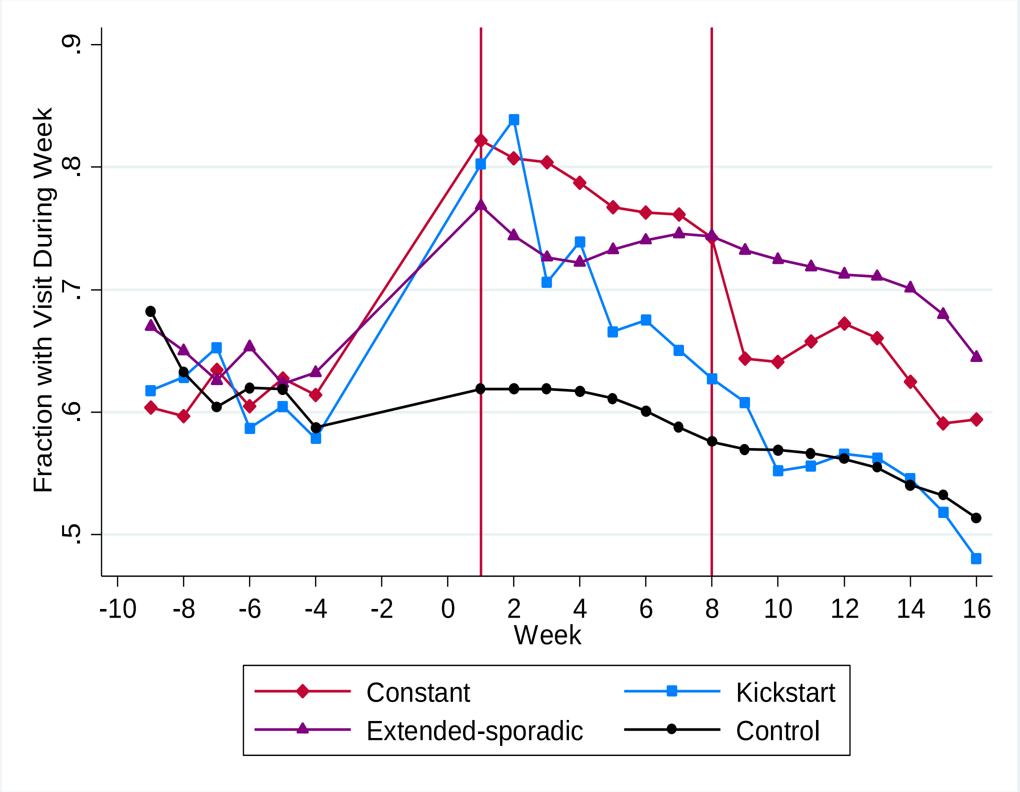

Figure notes: Figure presents LOESS smoothed plots with bandwidth set at 0.4. Outcome variable is an indicator for making at least one visit to the company gym during that week. Week 1 (marked with vertical line) denotes the first week of the incentive program. Pre-treatment visits rates for weeks -9 through -4 are included in the graph. Weeks -3 through 0 are omitted because these were weeks in which the initial survey and/or information about the treatment assignment was sent to subjects.

Table 1. Treatments

Treatment	Sample	Length of Incentive Period	Length of Program	Maximum Earnings	Description of Incentives
Constant	Members & Non-Members	8 weeks	8 weeks	$160	$10/visit for 8 weeks (capped at 2 visits/week)
Kickstart	Members & Non-Members	8 weeks	8 weeks	$160	$25/visit for 1st 2 weeks of incentive period; $5/visit the last 6 weeks of incentive period (capped at 2 visits/week)
Constant Short	Non-Members	4 weeks	4 weeks	$80	$10/visit for 4 weeks (capped at 2 visits/week)
Extended-Sporadic	Members	8 weeks	16 weeks	$160	$10/visit for a random 8 weeks out of a 16 week period (capped at 2 visits/week)
Control	Members & Non-Members	NA	NA	$0	NA

Table 2. Summary Statistics for Non-Member and Member Participants

Panel A. Non-members

Variable	Overall Mean	Control Mean	Constant Difference	Kickstart Difference	Constant-Short Difference	P-value of All Treatments=0
Age	41.29	41.24	-0.25	-0.22	-0.28	1.00
	[11.43]	[12.1]				
Male	0.52	0.49	0.05	0.05	0.06	0.84
College or More	0.7	0.7	0.03	-0.02	0.03	0.8
Interest in Joining Gym	-0.83	-0.77	0.07	0.06	-0.02	0.94
(-2 to 2, most interested)	[1.35]	[1.32]				
Married	0.67	0.69	-0.01	-0.03	-0.04	0.89
One or More Children at Home	0.5	0.48	0	0.08	-0.03	0.3
Days of Exercise Last Week	2.24	2.31	-0.02	-0.28	-0.17	0.58
	[1.81]	[1.75]				
Target Days of Exercise	3.6	3.59	0.01	-0.1	-0.16	0.77
	[1.54]	[1.38]				
Subjective Probability of Meeting Target	58.91	61.61	0.59	-6.76	-3.29	0.23
	[32.4]	[30.22]				
Days Exercised at Different Gym in Week	0.6	0.65	-0.05	-0.03	0.03	0.97
	[1.35]	[1.35]				
Days Exercised not at Gym in Week	1.85	1.82	0.08	0.01	-0.14	0.80
	[1.79]	[1.65]				
Body Mass Index	28.16	27.57	1.04	0.66	1.09	0.54
	[6.18]	[5.74]				
Fraction Overweight or Obese	0.66	0.61	0.13	0.06	0.03	0.18
Number of Observations	609	109	180	182	138	

Table notes: The overall mean column is the overall mean. The control mean column is the control group mean. The constant difference column is the mean difference between the constant group and the control group; they come from regressions that include cohort fixed effects. The kickstart difference column is the mean difference between the kickstart group and the control group; they come from regressions that include cohort fixed effects. The constant-short difference column is the mean difference between the constant-short group and the control group; they come from regressions that include cohort fixed effects. The p-value column displays the p-values testing that the means of all 4 groups (3 treatment groups + 1 control) are equal. For the non-dichotomous variables, the numbers in brackets represent the standard deviations.

Table 2. Summary Statistics for Non-Member and Member Participants

Panel B. Members

					Extended	
	Overall	**Control**	**Constant**	**Kickstart**	**Sporadic**	**P-value of All**
Variable	**Mean**	**Mean**	**Difference**	**Difference**	**Difference**	**Treatments=0**
Age	40.72	41.39	-1.56	-1.16	0.18	0.61
	[10.46]	[11.43]				
Male	0.48	0.42	0.07	0.11	0.04	0.55
College or More	0.72	0.79	-0.09	-0.04	-0.12	0.36
Married	0.71	0.75	0.01	-0.15	-0.01	0.05
One or More Children at Home	0.56	0.59	-0.03	-0.01	-0.03	0.98
Days of Exercise Last Week	3.57	3.79	-0.31	-0.24	-0.23	0.75
	[1.81]	[1.81]				
Target Days of Exercise	4.52	4.67	-0.13	-0.24	-0.19	0.61
	[1.15]	[1.14]				
Subjective Probability of Meeting Target	74.29	74.05	0.73	-0.05	1.42	0.97
	[24.98]	[25.65]				
Days Exercised at Company Gym in Week	2.3	2.73	-0.61	-0.57	-0.35	0.20
	[1.84]	[1.93]				
Days Exercised at Different Gym in Week	0.43	0.55	-0.16	-0.2	-0.05	0.56
	[.97]	[1.12]				
Days Exercised not at Gym in Week	1.57	1.26	0.34	0.43	0.35	0.32
	[1.46]	[1.39]				
Body Mass Index	27.32	27.01	0.27	-0.5	1.41	0.48
	[7.77]	[6.22]				
Fraction Overweight or Obese	0.62	0.53	0.06	0.15	0.12	0.21
Number of Observations	371	63	100	102	106	

Table notes: The overall mean column is the overall mean. The control mean column is the control group mean. The constant difference column is the mean difference between the constant group and the control group; they come from regressions that include cohort fixed effects. The kickstart difference column is the mean difference between the kickstart group and the control group; they come from regressions that include cohort fixed effects. The constant-short difference column is the mean difference between the constant-short group and the control group; they come from regressions that include cohort fixed effects. The p-value column displays the p-values testing that the means of all 4 groups (3 treatment groups + 1 control) are equal. For the non-dichotomous variables, the numbers in brackets represent the standard deviations.

Table 3. Extensive Margin: Non-Members' Participation Decision

	(1)	(2)	(3)
Dependent variable:	**Joined**	**Joined**	**Joined**
Subgroup (if any):	***All***	***Low exercisers***	***High exercisers***
Constant-Short	0.032	-0.0001	0.11*
	(0.033)	(0.046)	(0.055)
Constant	0.10***	0.13**	0.13***
	(0.034)	(0.053)	(0.045)
Kickstart	0.10***	0.12**	0.11**
	(0.035)	(0.053)	(0.043)
Observations	596	318	253
Mean of dep var in control group	0.046	0.054	0

Table notes: Each column represents a separate regression. Robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1 "Low exercisers" are non-members reporting 2 or fewer days of exercise in the week prior to the initial survey as well as in their typical week. "High exercisers" are non-members reporting 3 or more days of exercise in the week prior to the initial survey as well as in their typical week. We drop from columns (2) and (3) the 25 subjects whose prior week exercise was at odds with their typical exercise.

Table 4. Treatment Effects among Non-Members

	(1)	(2)
Dependent variable:	**Any Visit**	**Number of Visits**
Weeks 1-4 * Constant	0.116***	0.275***
	(0.027)	(0.064)
Weeks 1-4 * Kickstart	0.113***	0.248***
	(0.027)	(0.067)
Weeks 1-4 * Constant-Short	0.064**	0.167**
	(0.026)	(0.067)
Weeks 5-8	0.002	-0.016
	(0.005)	(0.019)
Weeks 5-8 * Constant	0.106***	0.265***
	(0.025)	(0.057)
Weeks 5-8 * Kickstart	0.074***	0.155***
	(0.023)	(0.049)
Weeks 5-8 * Constant-Short	0.027	0.109**
	(0.020)	(0.054)
Weeks 9-16	0.007	0.001
	(0.013)	(0.030)
Weeks 9-16 * Constant	0.036**	0.104**
	(0.018)	(0.047)
Weeks 9-16 * Kickstart	0.030	0.072*
	(0.018)	(0.043)
Weeks 9-16 * Constant-Short	0.017	0.070
	(0.018)	(0.045)
Observations	9,536	9,536
Number of persons (cluster level)	596	596
Control pre period mean of dep var	0	0

Table notes: All regressions control for male, education degree fixed effects, age, interest in the company gym fixed effects, whether or not the individual is married, and cohort fixed effects. Each column represents a separate regression. Robust standard errors clustered on individual are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 5. Treatment Effects among Members

	(1)	(2)	(3)	(4)
			Number of	Number of
Dependent Variable:	**Any Visit**	**Any Visit**	**Visits**	**Visits**
Weeks 1-8	-0.017	-0.017	-0.091	-0.091
	(0.035)	(0.035)	(0.093)	(0.093)
Weeks 1-8 * Constant	0.184***	0.184***	0.433***	0.433***
	(0.048)	(0.048)	(0.135)	(0.135)
Weeks 1-8 * Kickstart	0.119***	0.119***	0.436***	0.436***
	(0.044)	(0.044)	(0.121)	(0.121)
Weeks 1-8 * Extended Sporadic	0.111**		0.434***	
	(0.044)		(0.134)	
Weeks 1-8 * Incentive ON for Extended Sporadic		0.158***		0.562***
		(0.045)		(0.137)
Weeks 1-8 * Incentive OFF for Extended Sporadic		0.065		0.311**
		(0.046)		(0.141)
Weeks 9-16	-0.073**	-0.073*	-0.304**	-0.304**
	(0.036)	(0.037)	(0.123)	(0.126)
Weeks 9-16 * Constant	0.093**	0.093*	0.233	0.233
	(0.047)	(0.048)	(0.157)	(0.161)
Weeks 9-16 * Kickstart	0.009	0.009	0.198	0.198
	(0.043)	(0.044)	(0.140)	(0.144)
Weeks 9-16 * Extended Sporadic	0.137***		0.491***	
	(0.046)		(0.161)	
Weeks 9-16 * Incentive ON for Extended Sporadic		0.165***		0.598***
		(0.048)		(0.169)
Weeks 9-16 * Incentive OFF for Extended Sporadic		0.109**		0.382**
		(0.047)		(0.163)
Observations	8,184	8,184	8,184	8,184
Number of persons (cluster level)	372	372	372	372
Control pre-period mean of dep var	0.624	0.624	1.897	1.897

Table notes: Each column represents a separate regression. All regressions include individual fixed effects. "Incentive ON for Extended Sporadic" represents the weeks in which members of the extended sporadic group were incentivized to visit. These 8 weeks were randomly allocated within weeks 1-16, separtely for each participant. Robust standard errors clustered on individual are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 6. Treatment Effects among Members by Ex-Ante Attendance Rates

	(1)	(2)	(3)	(4)
Dependent variable:	**Any Visit**	**Any Visit**	**Number of Visits**	**Number of Visits**
Subgroup:	***Less than 2 visits in pre period***	***2 or more visits in pre period***	***Less than 2 visits in pre period***	***2 or more visits in pre period***
Weeks 1-8	0.045	-0.058	0.108	-0.223*
	(0.067)	(0.037)	(0.116)	(0.130)
Weeks 1-8 * Constant	0.325***	0.080*	0.774***	0.180
	(0.089)	(0.044)	(0.189)	(0.166)
Weeks 1-8 * Kickstart	0.172**	0.062	0.485***	0.354**
	(0.082)	(0.043)	(0.168)	(0.160)
Weeks 1-8 * Extended Sporadic	0.217**	0.054	0.692***	0.300*
	(0.085)	(0.044)	(0.208)	(0.164)
Weeks 9-16	-0.005	-0.117***	0.068	-0.548***
	(0.059)	(0.045)	(0.119)	(0.178)
Weeks 9-16 * Constant	0.116	0.074	0.293	0.170
	(0.081)	(0.053)	(0.182)	(0.215)
Weeks 9-16 * Kickstart	-0.038	0.037	-0.082	0.364*
	(0.070)	(0.054)	(0.136)	(0.209)
Weeks 9-16 * Extended Sporadic	0.235***	0.086	0.729***	0.381*
	(0.078)	(0.053)	(0.213)	(0.207)
Observations	3,366	4,818	3,366	4,818
Number of persons (cluster level)	153	219	153	219
Control pre period mean of dep var	0.220	0.890	0.387	2.890

Table notes: Each column represents a separate regression. All regressions include individual fixed effects. Robust standard errors clustered on individual are in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Members visiting less than 2 visits in pre-period are those who in weeks -4 to -9 went to the gym 2 or more days fewer than half of the weeks. Members visiting 2 or more visits in pre-period are the remaining members.

Table 7. Cost Effectiveness by Gym Membership and Treatment Status

	Members			Non-members (with Gym Member Bonus)		
	Constant	Kickstart	Sporadic	Constant	Kickstart	Constant Short
Visits induced per dollar spent	0.04	0.05	0.08	0.15	0.11	0.21
Dollars spent per visit induced	$22.76	$19.42	$12.74	$6.70	$9.21	$4.77

Table Notes: These calculations are based on the treatment effect on number of visits divided by the per-subject cost of the incentive program. The treatment effects are computed over weeks 1-24.