Management Science

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

Creating Exercise Habits Using Incentives: The Trade-off Between Flexibility and Routinization

John Beshears, Hae Nim Lee, Katherine L. Milkman, Robert Mislavsky, Jessica Wisdom

To cite this article:
John Beshears, Hae Nim Lee, Katherine L. Milkman, Robert Mislavsky, Jessica Wisdom (2020) Creating Exercise Habits Using Incentives: The Trade-off Between Flexibility and Routinization. Management Science

Published online in Articles in Advance 15 Oct 2020

. https://doi.org/10.1287/mnsc.2020.3706

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

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–33
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Creating Exercise Habits Using Incentives: The Trade-off Between Flexibility and Routinization

John Beshears,[a] Hae Nim Lee,[b] Katherine L. Milkman,[b] Robert Mislavsky,[c] Jessica Wisdom[d]

[a] Harvard Business School, Boston, Massachusetts 02163; [b] The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [c] Carey Business School, Johns Hopkins University, Baltimore, Maryland 21202; [d] Google, Mountain View, California 94043

Contact: jbeshears@hbs.edu, http://orcid.org/0000-0002-3808-6423 (JB); leehn@wharton.upenn.edu (HNL); kmilkman@wharton.upenn.edu, http://orcid.org/0000-0002-9706-4830 (KLM); mislavsky@jhu.edu, http://orcid.org/0000-0002-9620-3528 (RM); jessie@humu.com (JW)

Received: March 7, 2019
Revised: November 27, 2019; May 4, 2020
Accepted: May 8, 2020
Published Online in Articles in Advance: October 15, 2020

https://doi.org/10.1287/mnsc.2020.3706

Abstract. Habits involve regular, cue-triggered routines. In a field experiment, we tested whether incentivizing exercise routines—paying participants each time they visit the gym within a planned, daily two-hour window—leads to more persistent exercise than offering flexible incentives—paying participants each day they visit the gym, regardless of timing. Routine incentives generated *fewer* gym visits than flexible incentives, both during our intervention and after incentives were removed. Even among subgroups that were experimentally induced to exercise at similar rates during our intervention, recipients of routine incentives exhibited a larger decrease in exercise after the intervention than recipients of flexible incentives.

History: Accepted by Yan Chen, decision analysis.
Funding: We acknowledge financial support from the Marketing Science Institute [Grant 4-1916]; Google; the National Institute on Aging [Grants P01AG005842, P30AG034532, and P30AG034546]; Harvard Business School; and Wharton School, University of Pennsylvania.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2020.3706.

Keywords: economics • behavior and behavioral decision making • healthcare • exercise • habit • routine

1. Introduction

Small, repeated, everyday decisions can have profound effects on many critical life outcomes. Choices that may seem trivial in the moment, such as how much to exercise, what to eat, how hard to study, and how to spend money, often accumulate over time to have large consequences (e.g., Mokdad et al. 2004, Kuh et al. 2006, Schroeder 2007). Interventions capable of shifting the habits that govern many everyday behaviors could improve individual welfare tremendously if applied to decisions about health, education, and personal finance (e.g., Beshears et al. 2013, Gertler et al. 2014, Loewenstein et al. 2016).[1] Companies, recognizing the importance of such behaviors for employee well-being and productivity, are increasingly interested in promoting positive employee habits in these domains. For example, more than 90% of employers with at least 200 employees offer workplace wellness programs, and 63% of employers with wellness offerings sponsor a program that encourages exercise habits (Mattke et al. 2012, Jones et al. 2019).

Psychology research has shown that stable habits tend to be characterized by engagement in behaviors under consistent circumstances or "routine" conditions; they are typically done at the same time, in the same place, and following the same cue to act (Wood and Neal 2016, Wood and Rünger 2016). As an individual develops a pattern of repeatedly responding to a given configuration of contextual cues by rehearsing a specific set of behaviors, that configuration of cues gradually begins to trigger a mental representation of the behavioral response without requiring the exertion of executive control. This fluid and effortless mental association makes performance of the habitual behavior largely automatic (Wood and Rünger 2016). The exact neural mechanisms by which this process occurs are still debated, but the nature of a habit can be conceptualized as a reduction in the cognitive effort associated with engaging in the habitual behavior when routine contextual cues are in place. These patterns suggest an opportunity for organizations: routines are central to habitual behavior, and organizations may be able to capitalize on this fact when attempting to encourage the formation of beneficial habits.

Past research evaluating interventions that encourage routines has shown promising results (Lally et al. 2008; Carels et al. 2011, 2014; Judah et al. 2013; for a discussion, see Wood and Rünger 2016). However, prior interventions designed to facilitate habit formation have done far more than simply encouraging routines. For example, they have provided general lifestyle advice (Lally et al. 2008; Carels et al. 2011, 2014) or coupled one behavior (e.g., flossing) with another (e.g., toothbrushing) (Judah et al. 2013).

Furthermore, previous studies attempting to encourage routines were small in scale (each with sample sizes of approximately 100 participants or fewer). To address this gap in the literature, we conducted a 2,508-participant field experiment designed to test whether people form longer-lasting exercise habits if they are encouraged to maintain a strict routine rather than encouraged to exercise frequently without necessarily adhering to a particular schedule.

Our field experiment included employees at Google who were interested in exercising more regularly at workplace gyms. At the beginning of the experiment, all participants chose a daily, two-hour window when it would be best for them to exercise, and all participants were informed that they would receive reminders to exercise every weekday at the beginning of that window. Participants were then randomly assigned to one of five experimental conditions. During the four-week intervention, participants in two "flexible" experimental conditions were paid $3 and $7, respectively, for any weekday when they exercised for at least 30 minutes at a workplace gym. Participants in two "routine" experimental conditions were also paid $3 and $7 for these workouts but *only* if they entered the gym within their chosen two-hour window. Participants in the control group received no monetary incentives for exercise. We analyzed data on participants' gym visits both during the four-week intervention period and after the intervention period, when incentive payments were no longer offered.

During the intervention period, the two flexible conditions (pooled together) increased the number of gym visits by 0.19 per week relative to the two routine conditions (pooled together), and they increased the likelihood of having at least one gym visit in a given week by 4 percentage points relative to the routine conditions. Of course, for the purpose of learning about habit formation, we are even more interested in the persistence of treatment effects after the intervention period ended and the financial incentives were removed. During the first four weeks of the postintervention period, the pooled flexible conditions increased the number of gym visits by 0.10 per week and increased the likelihood of having at least one gym visit in a given week by 6 percentage points relative to the pooled routine conditions, although the former difference is only marginally statistically significant.

Although measuring the relative impact of the flexible and routine incentive schemes is relevant for judging the efficacy of these two types of policies, the difference in postintervention exercise we detect is likely driven by the fact that the flexible conditions produced more exercise than the routine conditions during the intervention (at greater expense to the employer). This finding that more exercise in the past begets more exercise in the future (regardless of the timing of past gym visits) is consistent with past empirical research (e.g., Charness and Gneezy 2009) and models of habit formation (e.g., Becker and Murphy 1988). Our study design—which randomized not only whether incentives for exercise were flexible or routine but also, their magnitude—allows us to ask a more interesting and novel question. Specifically, we can examine which type of incentive scheme generated more postintervention gym visits, holding constant the frequency of intervention-period gym visits. It turns out that the routine condition offering $7 incentive payments and the flexible condition offering $3 incentive payments generated approximately the same number of intervention-period gym visits, with the routine condition generating more in-window gym visits (visits that began during a participant's chosen two-hour window) and fewer out-of-window gym visits (visits that began outside a participant's chosen two-hour window). Past research on habits suggests that the routine condition participants who were offered $7 incentive payments should be more likely to develop exercise routines and should therefore sustain more persistent exercise habits after the intervention than the flexible condition participants who were offered $3 incentive payments.[2] If anything, however, we find the opposite result in our experimental data. In the transition from the intervention period to the first four weeks of the postintervention period, the routine condition offering $7 incentive payments exhibited a decrease in average weekly gym visits that was 0.14 larger (i.e., more negative) than the decrease observed in the flexible condition offering $3 incentive payments. A similar pattern emerged when examining the average weekly likelihood of at least one gym visit. Specifically, the decrease in the average weekly likelihood of at least one gym visit was five percentage points larger for the routine condition with $7 payments than for the flexible condition with $3 payments. These findings are reinforced by an instrumental variables analysis of our data, which leads to similar conclusions. In short, when people are induced to exercise at an equal frequency but in a more routinized way, we find evidence that they form *weaker* exercise habits, contrary to past theorizing.

To interpret our experimental results within a broader context, we present a simple model of habit formation, which could apply to visiting the gym, giving feedback to employees, or engaging in any other behavior that might be repeated in a consistent fashion. The model has a single agent and two periods. In each period, the agent has three possible decisions: taking an in-window action (that is, an action at a planned and consistent time), an out-of-window action (that is, an action at any other time within the period), or no action. Mapping the model to our experiment, the agent can have an in-window gym visit, an out-of-window gym visit, or no gym visit in a given period. The intrinsic utility of each type of action (in window or out of window, relative to no action) is randomly drawn at the beginning of each period.[3] In the first period, the agent may be offered financial incentives for taking an action. A routine incentive scheme offers a payment for an in-window action but not for an out-of-window action, whereas a flexible incentive scheme offers a payment for either action. An action of a given type in the first period forms a habit in the sense that it increases the intrinsic utility of that same type of action in the second period (an assumption that is consistent with our experimental results). In our baseline model, we do *not* assume that an in-window action in the first period has a stronger habit-forming effect than an out-of-window action in the first period. This assumption is easily accommodated within our framework and does not change our qualitative conclusions, but it is not necessary for the model to serve its purpose, which is to highlight other factors that influence the effectiveness of routine incentive schemes relative to flexible incentive schemes. In our analysis of the model, we take the perspective of a manager or policy maker who is fundamentally indifferent between in-window and out-of-window actions and who simply wishes to increase the overall likelihood that the agent takes either an in-window or an out-of-window action in the second period (i.e., in the long run).

With this setup, the model's predictions depend on the rates of in-window and out-of-window actions in the absence of incentives. For instance, if the likelihood of an in-window action in the absence of incentives is high and the likelihood of an out-of-window action in the absence of incentives is low, a routine incentive scheme leads to a larger increase in the overall likelihood of seeing either an in-window or an out-of-window action in the second period than a flexible incentive scheme offering payments of the same dollar amount. If, on the other hand, the likelihood of an in-window action is similar to or less than the likelihood of an out-of-window action in the absence of incentives, a flexible incentive scheme is more effective at increasing the overall likelihood of seeing either an in-window or an out-of-window action in the second period than a routine incentive scheme offering payments of the same dollar amount.[4] Our experimental data appear to match this latter case. Intuitively, if opportunities for routine (in-window) gym visits are often inferior to alternative (out-of-window) gym visit opportunities, an incentive scheme that promotes routine exercise tends to generate in-window gym visits that supplant out-of-window gym visits, instead of in-window gym visits that take place when no gym visit would have otherwise occurred. In our data, an increase in in-window gym visits appears to support habit formation, but a reduction in out-of-window gym visits simultaneously appears to undermine other routines that might have developed. Our field experiment suggests that this problem can arise in dynamic, fast-paced workplaces, where it is difficult to identify a regular time window for exercise that is unlikely to be disrupted or superseded by alternative exercise opportunities. We conclude that although routines have been proven elsewhere to be important to habit formation, it may be challenging for managers to encourage routines in environments with frequently shifting demands on people's time. Routine incentive schemes may be more effective when applied to behaviors and environments for which the best opportunity to take the desired action is consistent from one time period to the next. In such stable contexts, routine incentive schemes can potentially strengthen incipient habits.

This paper is related to several strands of prior research. First, we build on previous applications of psychological insights to design interventions that change behavior (Madrian and Shea 2001, Johnson and Goldstein 2003, Thaler and Benartzi 2004, Larrick and Soll 2008, Thaler and Sunstein 2008, Benartzi et al. 2017). In particular, recent research has shown that interventions rewarding repeated engagement in desirable behaviors like exercise, for as little as a month, can build habits that stay in place after incentives are removed (Charness and Gneezy 2009, Acland and Levy 2015, Royer et al. 2015, Hussam et al. 2017). These findings are consistent with prior work theorizing that habits are formed by repeatedly engaging in a behavior (Becker and Murphy 1988). Our paper extends this line of work by testing an intervention that leverages psychological insights about the importance of repeated engagement in a behavior *in a routine fashion* for the purposes of forming a habit.

The idea that routines are important for habit formation (Wood and Neal 2016, Wood and Rünger 2016) is based in part on experimental studies of habitual behaviors. For example, individuals with a strong prior habit of eating popcorn in movie theaters ate the same amount of popcorn whether it was fresh or stale if they were sitting in a movie theater but ate more fresh popcorn than stale popcorn if they were sitting in a meeting room, suggesting that automatic performance of a habitual behavior (eating popcorn, regardless of whether it is fresh or stale) is associated with routine conditions (sitting in a movie theater) but not with nonroutine conditions (sitting in a meeting room). Individuals without a strong prior habit of eating popcorn in movie theaters ate more fresh popcorn than stale popcorn both in the movie theater and in the meeting room (Neal et al. 2011).

There is also previous research that directly studies individuals who have succeeded in establishing beneficial habits. In the domain of medication regimens, adherence is higher among those with regular pill-taking routines (Brooks et al. 2014). In a sample of regular gym visitors, 75% reported that they tended to exercise at the same time of day, and although exercising at the same time of day did not correlate with exercise frequency within this highly selected sample, the large fraction of individuals in the sample who indicated that time of day was part of their exercise routine suggests that consistent timing may be helpful for forming a habit (Tappe et al. 2013). As described, routine-building interventions that have been evaluated in previous research have used small sample sizes and have either incorporated (a) more features than the mere encouragement of routines or (b) context-specific design elements that make generalization difficult (Lally et al. 2008; Carels et al. 2011, 2014; Judah et al. 2013). This paper reports the results of a larger-scale field experiment focused on daily routines that could be a broadly applicable path to promoting beneficial habits.

The remainder of this paper is organized as follows. Section 2 presents our experimental design and our methods for analyzing the data. Section 3 presents our experimental results. In Section 4, we analyze the simple model that we use to interpret our findings, and we discuss the limitations of our study. Section 5 concludes.

2. Experimental Design and Implementation

2.1. Setting

We collaborated with the technology company Google to conduct a randomized controlled trial with a subset of the company's employees. To be eligible to participate, an individual was required to be a full-time, part-time, or fixed-term employee or an intern at one of the company's seven U.S. office locations that partnered on our study, leaving us with roughly 25,000 eligible employees. Each office location where our experiment was implemented had at least one on-site fitness center. Although each fitness center boasts unique features, all offer personal trainers and group fitness classes, and all are equipped with exercise machines and weights. Basic gym access (e.g., use of the exercise machines and weights) is free to all employees, but employees must pay fees for extra services, such as personal training, nutrition counseling, and some special group classes. Upon entering the gym, employees encounter a computer kiosk where they are asked to swipe their employee identification badge and record their gym visit. We rely on these log-in data to track individual gym attendance. Employees are also asked to swipe their badge as they exit the gym.

2.2. Participant Recruitment and Randomization

Figure 1 shows the flow and randomization of study participants, and Figure 1 in the online appendix illustrates the timeline of the experiment.

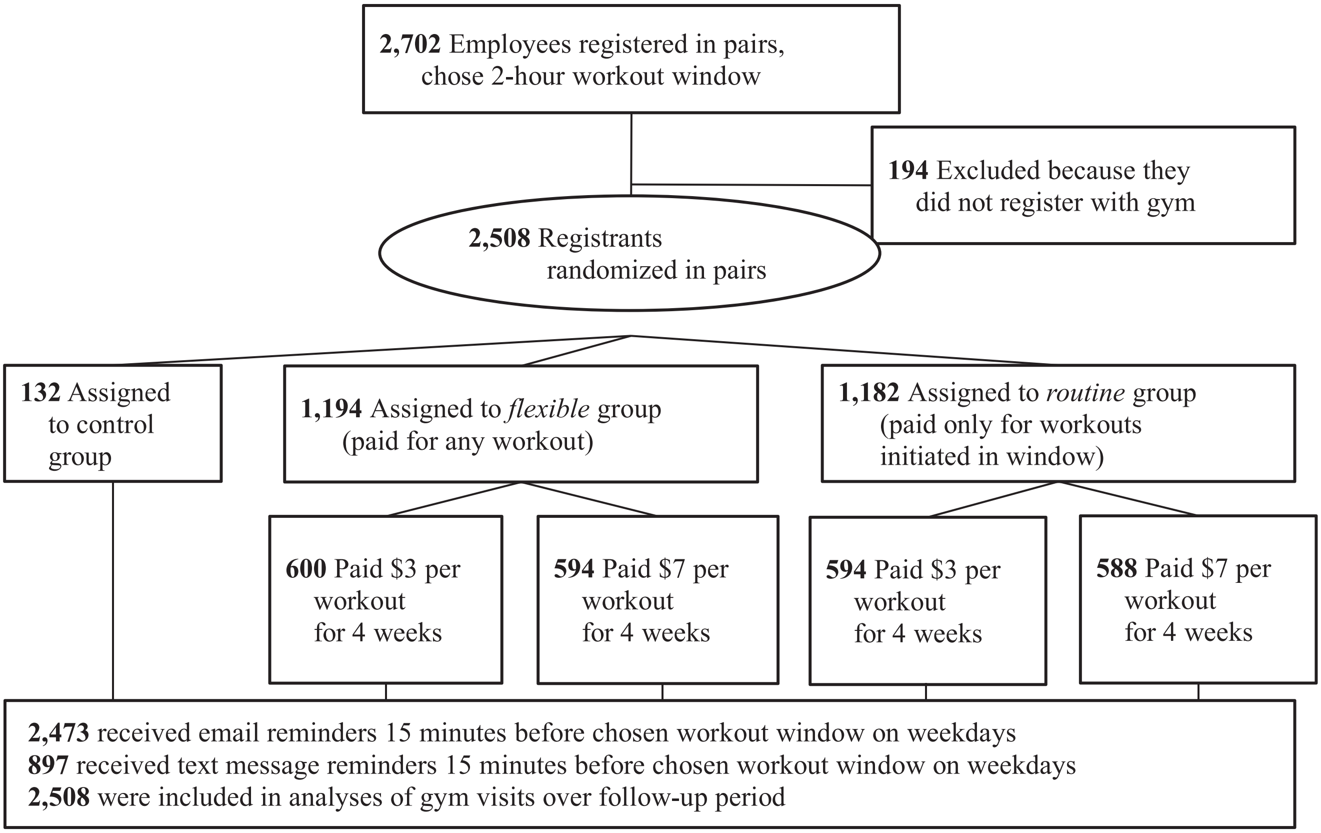
2.2.1. Recruitment. Participant recruitment began on February 3, 2015, through a series of poster and email advertisements (see Online Appendix B, Figures B1 and B2). These advertisements explained that employees had a chance to be paid for exercising and encouraged employees to visit an internal company website to learn more and register with a friend from their office by February 23, 2015 (a deadline that was subsequently extended by two days to accommodate additional recruiting efforts). The posters and emails informed employees that completing an initial registration survey would enter them into raffles for a Fitbit Surge (a fitness tracker valued at approximately $250) and a $100 entertainment gift card.

2.2.2. Registration Survey. Google employees who responded to our recruitment campaign were given a web link to complete our registration survey (see Online Appendix B, Figure B3). Upon starting the survey, employees were told that the program, labeled the Fresh Start Fitness Challenge, was part of a research study being conducted by Google in partnership with academic researchers and was designed to help employees achieve their fitness goals. They were also reminded about the raffles and were told that completion of the survey did not guarantee registration in the study in the event of overenrollment.

The survey began with a consent form and some background questions (name, email address, office location, typical number of days per week involving exercise for at least 30 minutes, gender, and ethnicity). Next, employees were asked to register their employee identification badge with the Google gym, allowing us to track their gym entrances and exits (see Online Appendix C for additional details about the gym registration process). After being prompted to register with the gym, participants were asked to select a "workout buddy" (their partner for the program) by providing the name and corporate email address of another employee at the same office location. This employee then received an email with a prompt to complete the registration survey (see Online Appendix D for more detailed information about the partner pairing process).

After choosing a workout partner, employees were asked to select a two-hour block of time when they preferred to start their weekday workouts (which would last at least 30 minutes) at the company gym.[5]

Figure 1. Experimental Flowchart


2,702 Employees registered in pairs, chose 2-hour workout window
194 Excluded because they did not register with gym
2,508 Registrants randomized in pairs
132 Assigned to control group
1,194 Assigned to *flexible* group (paid for any workout)
1,182 Assigned to *routine* group (paid only for workouts initiated in window)
600 Paid $3 per workout for 4 weeks
594 Paid $7 per workout for 4 weeks
594 Paid $3 per workout for 4 weeks
588 Paid $7 per workout for 4 weeks
2,473 received email reminders 15 minutes before chosen workout window on weekdays
897 received text message reminders 15 minutes before chosen workout window on weekdays
2,508 were included in analyses of gym visits over follow-up period

Based on informal conversations with Google employees, we made the workout windows two-hours long. Our goal was to strike a reasonable balance between windows that would (a) sufficiently accommodate day-to-day variability in employees' schedules and (b) be sufficiently narrow to ensure a series of gym visits initiated at different times within a window would still constitute a time-based routine. Although employees could coordinate workout windows with their partners (31.6% of the final sample selected a workout window that overlapped perfectly[6] with their partner's), they were not required to do so. Employees were then told that they would receive daily reminders (sent to their corporate email address) Monday through Friday, 15 minutes prior to the start of their workout window. They could also opt in to receive text message reminders at the same time by providing their cell phone number (35.8% of the final sample received text message reminders).

At this point in the registration survey, employees were offered a $10 Amazon gift card to create an (optional) account with AchieveMint, a free app that aggregates data from other apps and fitness trackers, including minute-by-minute step data from Fitbit, which we would collect for this study. Among the employees who were enrolled in the study, 25.9% (650 individuals) created an AchieveMint account and received a $10 gift card, and 4.5% (114 individuals) synched a Fitbit with AchieveMint.

Employees were then told that they were officially registered for the study and received a confirmation email (see Online Appendix B, Figure B4). At this point, participants could exit the survey or continue to optional demographic questions (e.g., age, height, weight, employment information, and current exercise habits).[7] Of the employees who were enrolled in the study, 54% completed all of these optional questions.

In total, 2,508 employees, or approximately 10% of the eligible population, successfully completed all steps of the registration process for our study.

2.2.3. Experimental Conditions.

Each participating pair of employees was randomly assigned to one of five conditions (four treatment conditions and one control condition). Participants in the control condition did not receive monetary payments for completing workouts. Participants in the treatment conditions received monetary payments when they completed a qualifying workout during the four-week intervention period. Two of the treatment conditions were *flexible* conditions, in which participants earned a payment for each weekday (Monday to Friday) during which they worked out at the company gym for at least 30 minutes. The other two treatment conditions were *routine* conditions, in which participants earned a payment for each weekday during which they worked out at the company gym for at least 30 minutes *provided that* they started the workout during their preselected workout window. For both the *flexible* and *routine* conditions, participants were randomly assigned to receive either $3 per workout or $7 per workout. In summary, the five experimental conditions were the *control* group, the *flexible $3* payment group, the *flexible $7* payment group, the *routine $3* payment group, and the *routine $7* payment group. It is worth noting that we randomized incentive size as well as the presence of *routine* versus *flexible* incentives because we anticipated that the *routine* and *flexible* conditions would induce different numbers of gym visits during the intervention period if incentives per qualifying gym visit were equal across conditions. By varying incentive size, we hoped to make it possible to compare the effects of routine versus flexible exercise postintervention given roughly equal exercise levels during the intervention.

2.2.4. Power Calculations. At the outset of this experiment, it was unclear how many of the tens of thousands of employees recruited to participate in our exercise program would enroll. We used the following method to conduct power calculations and to determine how many experimental conditions it would be possible to include in our study. First, we consulted prior research on encouraging gym attendance in healthy populations to assess the typical size of the effect of financial incentives on an individual's number of gym visits per week (Charness and Gneezy 2009, Milkman et al. 2014, Acland and Levy 2015, Royer et al. 2015). We found that incentives of approximately the same magnitude as ours have increased the number of gym visits per week by 10%–200%, with a typical standard deviation of 1.25 visits per week. Because prior research has reliably shown a large and significant effect of incentives on subsequent exercise habits, we determined that we could replicate this well-established finding by using a holdout control group that was small relative to our treatment groups. To meet our goal of having 80% power to detect a 35% difference between our *control* group and our *flexible $3* payment group, we aimed to assign 135 participants to the *control* group (we ended up with 132 in the control). In addition, we aimed for 80% power to detect a 15%–20% difference in gym visits between the *flexible* and *routine* conditions, which required approximately 750 participants per condition.[8]

We hoped to include up to eight treatment arms in our study. In addition to the four treatment arms described previously, we planned to incorporate up to four additional treatment conditions, which would have been identical to the four included treatment conditions except that participants would have been required to coordinate their workout windows with their workout partners. The purpose of these coordinated conditions would have been to assess the effects of social support on the creation of exercise habits. We decided in advance that if fewer than 3,135 (=750 × 4 + 135) employees signed up for the study, we would only include the conditions that allowed participants to select their workout windows individually. We implemented this plan when 2,702 employees signed up for the study (2,508 of whom completed all of the steps necessary for registration). This explains why our recruitment materials and intake survey encouraged employees to sign up for the study with a workout partner.

2.2.5. Randomization. Our registration survey closed on February 25, 2015 (two days later than initially planned because we extended our registration deadline to allow for additional recruiting efforts), and participants were randomized in pairs into one of the five experimental conditions on three separate dates (February 26, February 27, and March 3) depending on when they fulfilled all requirements for randomization. In order to proceed to randomization, participants must have (a) been partnered successfully and (b) registered online with the company gym.[9] On February 26, 1,582 individuals (791 pairs) were randomized, followed by 826 additional individuals (413 pairs) on February 27 and 100 individuals (50 pairs) on March 3. In total, 2,508 participants (1,254 pairs) were randomly assigned to conditions.

For each of the three randomization waves, we used a stratified randomization procedure with four strata based on (a) whether the average of the two partners' self-reported typical number of workouts per week was above or below the median within their randomization wave (the median for all waves was 2.5 workouts per week) and (b) whether the partners had (spontaneously) coordinated their workout windows. The randomization scheme therefore had 12 strata total, 4 for each of the three randomization waves. All regression results that we report control for strata fixed effects.

2.3. The Intervention

2.3.1. Information Provided to Participants About Their Experimental Conditions. As soon as a participant was randomized to an experimental condition, he or she received an email containing a link to a website describing the incentive structure for his or her condition and to a comprehension check survey (see Online Appendix E for details regarding this process). To encourage participants to read the treatment information, they were truthfully told that they would learn the registration raffle results as well as more details about their incentives after they completed the survey. Participants were also asked not to speak to anyone other than their workout buddy about the Fresh Start Fitness Challenge. However, we could not monitor or enforce compliance with this request.

2.3.2. Intervention Period. The intervention period began on March 2, 2015, for participants who were randomized in February and on March 4, 2015, for participants who were randomized on March 3. The intervention period ended on March 31, 2015, for all participants. Participants in all five conditions received daily workout reminder emails and/or text messages 15 minutes before the start of their self-selected workout window (see Online Appendix B, Figure B13 for the exact contents of the reminder messages).[10]

2.3.3. Postintervention Period. To encourage participants to continue to reliably swipe their employee identification badges when entering and exiting the gym, we sent emails to participants on April 1 (the first day after the conclusion of the intervention period) with the following announcement, among others: "On a randomly selected day in the month of April, we will have a lottery to select several members of the Challenge to receive $250 each. Here's the catch: you can only win if you badged in and out at the gym on the day of the lottery" (see Online Appendix B, Figure B14 for the full text of the emails). We later announced that we would hold this lottery every month through the end of 2015.

On April 17, 2015 (two weeks after the intervention period ended), participants received an email (see Online Appendix B, Figure B15) asking them to complete an exit survey (see Online Appendix B, Figure B16). After the exit survey was completed, participants received study-related payments through an online payment system. During the postintervention period, we continued to collect gym attendance data. In addition, participants continued to receive daily workout reminders for 10 months postintervention (until February 1, 2016) unless they opted out.

2.4. Statistical Analysis

2.4.1. Dependent Variables. Our primary outcome was participant gym attendance. To measure gym attendance, we obtained data tracking each time a study participant used his or her employee identification badge to enter or exit a company gym. Consistent with previous studies (Charness and Gneezy 2009, Milkman et al. 2014, Acland and Levy 2015), we initially planned to obtain and analyze data from two postintervention follow-up periods: (1) the 4-week period following the conclusion of the intervention (a length of time mirroring the length of our intervention) and (2) the 10-week period following the conclusion of the intervention (mapping roughly onto the follow-up periods from Charness and Gneezy 2009, study 1; Milkman et al. 2014; and Acland and Levy 2015). However, we learned in the midst of implementing the study that we would also be able to obtain data through the end of the calendar year, which concluded 40 weeks after the end of the intervention period, and we therefore analyze these supplemental data in addition to the data we planned to collect. In the main text of this paper, we focus on analyses of the four-week postintervention period, but analogous analyses for postintervention weeks 5–10 and 11–40 can be found in Figures 4 and 5 and Tables 2–5 in the online appendix.

Following past research on the impact of incentives on gym attendance habits, we measure gym attendance in two ways. First, we measure the total number of days each of our study participants visited the gym in each week (e.g., Charness and Gneezy 2009, Milkman et al. 2014, Acland and Levy 2015, Royer et al. 2015). Second, we measure whether a participant visited the gym at least once in a given week (e.g., Royer et al. 2015). This second dependent variable is a binary variable that is coded as one if a participant visited the gym at least once during the week and zero otherwise. For both dependent variables, we count a gym visit as having occurred as long as we see a study participant badge in at the gym.[11]

We also separately measure "in-window gym visits" and "out-of-window gym visits." We calculate the total number of in-window gym visits as the total number of days in a given week on which a participant was recorded as having made a gym visit within his or her preselected two-hour workout window (e.g., between 1:00 and 3:00 p.m. if the participant chose 1:00–3:00 p.m. as his or her preferred workout window during the registration survey). Analogously, the total number of out-of-window gym visits is defined as the total number of days within a specific week on which a participant made a gym visit outside his or her prespecified exercise window, even if he or she made an in-window visit on the same day. Thus, it is possible that participants' sum of in-window and out-of-window gym visits exceeds their total weekly gym visits because the total weekly gym visits variable records at most one visit per day.[12]

2.4.2. Regression Specifications. Our primary regression specification is

$$\begin{aligned} y_{it} = \alpha_0 + \alpha_1 C_{Flex\$3,i} + \alpha_2 C_{Flex\$7,i} + \alpha_3 C_{Rout\$3,i} \\ + \alpha_4 C_{Rout\$7,i} + \boldsymbol{\beta}' \boldsymbol{X_i} + \varepsilon_i, \end{aligned}$$

where *i* indexes participants and *t* indexes weeks. The right-hand-side variables of interest are indicators for experimental conditions ($C_{Flex\$3,i}$, $C_{Flex\$7,i}$, $C_{Rout\$3,i}$, and $C_{Rout\$7,i}$), and $\boldsymbol{X_i}$ is a vector of control variables. The control variables in our primary analyses are indicators for the 12 strata in our randomization scheme, with one indicator omitted to avoid collinearity with the constant term. The strata were defined by (a) randomization date (February 26, February 27, or March 3), (b) whether the average of the two partners' self-reported typical number of workouts per week was above or below the median within their randomization wave, and (c) whether the partners had (spontaneously) coordinated their workout windows. We conduct separate regressions for the four weeks of the intervention period and for the first four weeks of the postintervention period, and we cluster standard errors at the participant pair level. The left-hand-side variable y_{it} is one of six outcomes:

1. Number of days with a gym visit for participant *i* during week *t*
2. Number of days with an in-window gym visit for participant *i* during week *t*
3. Number of days with an out-of-window gym visit for participant *i* during week *t*
4. Whether participant *i* visited the gym at all during week *t*
5. Whether participant *i* visited the gym during his or her workout window during week *t*
6. Whether participant *i* visited the gym outside of his or her workout window during week *t*

We also conduct an analysis that uses the same regression framework but switches the outcome variable to be the change from the four intervention weeks to the first four postintervention weeks in the mean of one of the six variables. For this analysis, the regression sample includes one observation per participant. These results complement the results comparing levels of postintervention gym attendance across experimental conditions.

3. Results

3.1. Sample Summary

Table 1 presents summary statistics for self-reported variables collected in our preintervention registration survey: company tenure, weekly preintervention workout frequency,[13] body mass index (BMI; calculated from self-reported height and weight), gender, job function, and ethnicity. This table shows the mean, standard deviation, and proportion of participants who responded to each question for all participants in our study (column (1)), as well as for

Table 1. Summary Statistics Describing Study Participants Overall and by Condition

			Flexible			Routine		
	Total	Control	*Overall*	*$3*	*$7*	*Overall*	*$3*	*$7*
Number of years with company	3.08	3.25	3.16	3.31	3.02	2.96	3.16	2.75
	(2.60)	(2.38)	(2.69)	(2.80)	(2.56)	(2.54)	(2.69)	(2.35)
Proportion that responded	69%	77%	72%	70%	73%	67%	68%	65%
Self-reported workouts per week (preintervention)	2.67	2.64	2.69	2.66	2.72	2.65	2.62	2.68
	(1.54)	(1.63)	(1.51)	(1.54)	(1.49)	(1.56)	(1.59)	(1.54)
Proportion that responded	93%	98%	94%	94%	93%	93%	93%	93%
BMI	24.81	24.36	24.82	24.83	24.82	24.84	24.75	24.94
	(4.34)	(4.09)	(4.56)	(4.73)	(4.39)	(4.12)	(4.23)	(4.00)
Proportion that responded	67%	73%	68%	68%	69%	64%	66%	62%
Proportion of males	55%	55%	53%	52%	54%	57%	58%	56%
Proportion that responded	95%	97%	95%	95%	94%	94%	94%	94%
Job function								
Technology	61%	65%	60%	59%	60%	61%	62%	60%
Global business organization	21%	13%	21%	24%	19%	21%	22%	20%
General & administrative	19%	22%	19%	17%	21%	18%	16%	20%
Proportion that responded	69%	75%	71%	70%	73%	66%	68%	64%
Ethnicity								
White	49%	52%	49%	49%	50%	49%	50%	48%
Black	3%	3%	3%	2%	3%	3%	3%	2%
Asian	36%	35%	34%	36%	33%	37%	37%	38%
Hispanic	5%	3%	6%	6%	6%	4%	4%	4%
Mixed or other	7%	7%	8%	7%	8%	7%	6%	8%
Proportion that responded	89%	91%	88%	89%	88%	89%	88%	89%
Sample size	2,508	132	1,194	600	594	1,182	594	588

Notes. This table summarizes key employee characteristics based on responses to questions included in the registration survey, which participants had the option to skip. Because responding to these questions was voluntary, we report the proportion of participants who responded to each question. Standard deviations for means are reported in parentheses. Percentages may not add up to 100% because of rounding.

participants in our *control* group (column (2)), *flexible* groups (columns (3)–(5)), and *routine* groups (columns (6)–(8)). Performing pairwise statistical tests to compare each demographic variable across experimental conditions, we find that 4 of the 60 possible comparisons feature a difference that is statistically significant at the 5% level, roughly the number that would be expected by chance. Thus, it appears that random assignment successfully achieved balance across conditions.

Table 1 in the online appendix summarizes participant engagement with various aspects of the Fresh Start Fitness Challenge. Participants were required to receive workout reminder emails as of the date of randomization, and only 1%–2% opted out of receiving these emails by the end of the intervention period. Reminder text messages were an optional feature of the program, and 42% of participants opted to receive text messages as of the date of randomization, with only 6% subsequently opting out of text reminders during the intervention (leaving 36% still receiving text messages at the end of the intervention period). As for participants' chosen workout windows, 22% of participants selected workout windows beginning between 3:00 and 8:45 a.m., 29% selected windows beginning between 9:00 a.m. and 2:45 p.m., 48% selected windows beginning between 3:00 and 8:45 p.m., and 1% selected windows beginning between 9:00 p.m. and 2:45 a.m. Nearly one-third of participants had workout windows that exactly matched their partners', a fraction that is statistically significantly different from the 4% that would be expected if participants had their chosen workout windows but were randomly assigned to pairs.[14] Participants were also imperfect at predicting the workout windows that would correspond to their most regular gym visits. We determined this by looking at the timing of gym visits by participants in the control condition and flexible treatment conditions who had at least one weekday gym visit during our study's four-week incentive period. (We do not look at participants in the routine treatment conditions because they had monetary incentives encouraging gym visits during their selected workout windows.) We see that the mean fraction of incentive-period weekday gym visits that began during a participant's chosen workout window was 51%. Further, 64% of individuals could have selected a counterfactual workout window that would have had more incentive-period weekday gym visits in it than the chosen workout window.

Approximately one-quarter of participants signed up for an AchieveMint account, although we only track minute-by-minute physical activity data for the 4.5% of participants who linked a Fitbit device to their account.

3.2. Treatment Effects During the Intervention Period

Figure 2 in the online appendix presents means of weekly overall, in-window, and out-of-window gym attendance by experimental condition over the course of our four-week intervention period. The patterns indicate that larger incentive payments yielded more exercise, whereas routine incentives yielded more in-window workouts but fewer overall workouts. Tables 2 and 3 present the results of regressions that confirm these patterns. Note that each of the three outcome variables counts at most one gym visit per day. The sum of the control group means for the in-window visits and out-of-window visits variables exceeds the control group mean of the overall visits variable because participants could have recorded both an in-window visit and an out-of-window visit on the same day.

3.2.1. Incentive Size. Higher incentive payments led to more exercise during the intervention period. As Table 2, column (4) shows, participants paid \$7 per qualifying gym visit went to the Google gym a regression-estimated 0.30 more times per week than those paid \$3 ($p < 0.001$) and 0.79 more times per week than those in the control group ($p < 0.001$). The difference of 0.49 visits per week between participants paid \$3 per qualifying gym visit and those in the control group was also statistically significant ($p < 0.001$). As Table 3, column (4) shows, participants paid \$7 per qualifying gym visit went to the Google gym one or more times per week at a regression-estimated 6-percentage point higher rate than participants paid \$3 ($p < 0.001$) and at a regression-estimated 20-percentage point higher rate than those in the control group ($p < 0.001$). The difference of 14 percentage points between participants paid \$3 per qualifying gym visit and those in the control group was also statistically significant ($p < 0.001$).

3.2.2. Flexible vs. Routine Incentives. Table 2, column (7) shows that participants in the *flexible* conditions visited the gym a regression-estimated 0.19 times more per week during the intervention period than participants in the *routine* conditions ($p < 0.01$), and as Table 3, column (7) shows, participants in the *flexible* conditions visited the gym one or more times during a week at a regression-estimated 4-percentage point higher rate than participants in the *routine* conditions ($p < 0.05$). The point estimates for these effects are more than half the size of the point estimates for the effects of a \$4 increase in payments for qualifying gym visits (that is, the differences induced by raising payments from \$3 to \$7).

As expected, participants in the *routine* conditions exercised significantly more during their workout

Table 2. Regressions Predicting Participants' Weekly Workouts During the Intervention Period

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Total workouts	Total in-window workouts	Total out-of-window workouts	Total workouts	Total in-window workouts	Total out-of-window workouts	Total workouts	Total in-window workouts	Total out-of-window workouts
Flexible payment $3	0.58*** (0.14)	0.32*** (0.09)	0.27** (0.10)						
Flexible payment $7	0.89*** (0.14)	0.43*** (0.10)	0.50*** (0.10)						
Routine payment $3	0.40** (0.14)	0.57*** (0.10)	−0.15 (0.09)						
Routine payment $7	0.69*** (0.14)	0.96*** (0.10)	−0.21* (0.09)						
$3 Interventions				0.49*** (0.13)	0.45*** (0.09)	0.06 (0.09)			
$7 Interventions				0.79*** (0.13)	0.69*** (0.09)	0.15^{+} (0.09)			
Flexible interventions							0.74*** (0.13)	0.37*** (0.09)	0.38*** (0.09)
Routine interventions							0.54*** (0.13)	0.77*** (0.09)	−0.18* (0.09)
Mean values of control group	1.11	0.59	0.59	1.11	0.59	0.59	1.11	0.59	0.59
Observations	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032
R^2	0.07	0.07	0.10	0.07	0.05	0.04	0.07	0.06	0.10
Wald test ($3 flexible – $7 flexible)									
Difference in coefficients	−0.31*** (0.09)	−0.11 (0.07)	−0.24*** (0.07)						
Wald test ($3 flexible – $3 routine)									
Difference in coefficients	0.18* (0.09)	−0.25*** (0.08)	0.41*** (0.05)						
Wald test ($3 flexible – $7 routine)									
Difference in coefficients	−0.11 (0.09)	−0.64*** (0.08)	0.47*** (0.05)						
Wald test ($7 flexible – $3 routine)									
Difference in coefficients	0.49*** (0.09)	-0.14^{+} (0.08)	0.65*** (0.06)						

Table 2. (Continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Total workouts	Total in-window workouts	Total out-of-window workouts	Total workouts	Total in-window workouts	Total out-of-window workouts	Total workouts	Total in-window workouts	Total out-of-window workouts
Wald test ($7 flexible – $7 routine)									
Difference in coefficients	0.20* (0.09)	–0.53*** (0.08)	0.71*** (0.05)						
Wald test ($3 routine – $7 routine)									
Difference in coefficients	–0.29** (0.09)	–0.39*** (0.09)	0.06 (0.04)						
Wald test ($3 – $7)									
Difference in coefficients				–0.30*** (0.06)	–0.25*** (0.06)	–0.09* (0.04)			
Wald test (flexible – routine)									
Difference in coefficients							0.19** (0.06)	–0.39*** (0.06)	0.56*** (0.04)

Notes. This table reports a series of ordinary least squares regressions predicting a study participant's weekly number of (a) overall workouts, (b) workouts initiated during his or her workout window, and (c) workouts initiated outside of his or her workout window during the four-week intervention period. In each column, we report the mean number of workouts completed by the control group within this period. The primary predictors included in these regressions are treatment status indicators, which indicate the size of the incentive offered for exercise ($3 vs. $7) and the flexibility of the workout schedule (flexible vs. routine). We report pairwise Wald tests to assess whether all paired regression coefficients reported differ significantly from each other. Standard errors clustered by workout buddy pair are in parentheses. The control variables in the regressions are indicators for randomization strata (12 strata: three randomization dates, crossed with whether workout window perfectly overlapped with partner's workout window, crossed with whether self-reported pair number of workouts per week was above median for randomization date), as well as an indicator for missing workout window.

$^{+}p<0.10$; $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$.

Table 3. Regressions Predicting Participants' Likelihood of Working Out Each Week During the Intervention Period

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Any workouts? (Y/N)	Any in-window workouts? (Y/N)	Any out-of-window workouts? (Y/N)	Any workouts? (Y/N)	Any in-window workouts? (Y/N)	Any out-of-window workouts? (Y/N)	Any workouts? (Y/N)	Any in-window workouts? (Y/N)	Any out-of-window workouts? (Y/N)
Flexible payment $3	0.15***	0.13***	0.12***						
	(0.04)	(0.04)	(0.04)						
Flexible payment $7	0.23***	0.17***	0.20***						
	(0.04)	(0.04)	(0.04)						
Routine payment $3	0.13**	0.20***	−0.04						
	(0.04)	(0.04)	(0.04)						
Routine payment $7	0.17***	0.30***	−0.07*						
	(0.04)	(0.04)	(0.04)						
$3 Interventions				0.14***	0.17***	0.04			
				(0.04)	(0.03)	(0.04)			
$7 Interventions				0.20***	0.23***	0.06$^+$			
				(0.04)	(0.03)	(0.04)			
Flexible interventions							0.19***	0.15***	0.16***
							(0.04)	(0.03)	(0.04)
Routine interventions							0.15***	0.25***	−0.06
							(0.04)	(0.03)	(0.03)
Mean values of control group	0.50	0.31	0.32	0.50	0.31	0.32	0.50	0.31	0.32
Observations	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032
R^2	0.05	0.06	0.08	0.05	0.05	0.03	0.05	0.05	0.08
Wald test ($3 flexible – $7 flexible)									
Difference in coefficients	−0.07**	−0.04	−0.07**						
	(0.02)	(0.02)	(0.02)						
Wald test ($3 flexible – $3 routine)									
Difference in coefficients	0.03	−0.07**	0.16***						
	(0.02)	(0.03)	(0.02)						
Wald test ($3 flexible – $7 routine)									
Difference in coefficients	−0.02	−0.17***	0.20***						
	(0.02)	(0.02)	(0.02)						
Wald test ($7 flexible – $3 routine)									
Difference in coefficients	0.10***	−0.03	0.24***						
	(0.02)	(0.03)	(0.02)						

Table 3. (Continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Any workouts? (Y/N)	Any in-window workouts? (Y/N)	Any out-of-window workouts? (Y/N)	Any workouts? (Y/N)	Any in-window workouts? (Y/N)	Any out-of-window workouts? (Y/N)	Any workouts? (Y/N)	Any in-window workouts? (Y/N)	Any out-of-window workouts? (Y/N)
Wald test ($7 flexible – $7 routine)									
Difference in coefficients	0.05* (0.02)	–0.13*** (0.02)	0.27*** (0.02)						
Wald test ($3 routine – $7 routine)									
Difference in coefficients	–0.04$^+$ (0.02)	–0.09*** (0.03)	0.03$^+$ (0.02)						
Wald test ($3 – $7)									
Difference in coefficients				–0.06*** (0.02)	–0.07*** (0.02)	–0.02 (0.02)			
Wald test (flexible – routine)									
Difference in coefficients							0.04* (0.02)	–0.10*** (0.02)	0.22*** (0.02)

Notes. This table reports a series of ordinary least squares regressions predicting a study participant's weekly likelihood of completing a (a) workout anytime, (b) workout initiated during his or her workout window, and (c) workout initiated outside of his or her workout window during the four-week intervention period. In each column, we report the mean weekly fraction of participants in the control group who completed a workout within this period. The primary predictors included in these regressions are treatment status indicators, which indicate the size of the incentive offered for exercise ($3 vs. $7) and the flexibility of the workout schedule (flexible vs. routine). We report pairwise Wald tests to assess whether all paired regression coefficients reported differ significantly from each other. Standard errors clustered by workout buddy pair are in parentheses. The control variables in the regressions are indicators for randomization strata (12 strata: three randomization dates, crossed with whether workout window perfectly overlapped with partner's workout window, crossed with whether self-reported pair number of workouts per week was above median for randomization date), as well as an indicator for missing workout window. N, no; Y, yes.

$^+p<0.10$; $^*p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$.

windows than did participants in the *flexible* conditions. Participants in the *routine* conditions completed 1.34 in-window workouts per week on average and at least 1 in-window workout in a given week 56.2% of the time, whereas participants in the *flexible* conditions completed 0.95 in-window workouts per week on average and at least 1 in-window workout in a given week 46.2% of the time. Tables 2, column (8) and 3, column (8) show that the regression-estimated differences comparing these two outcome variables for the *routine* conditions versus the *flexible* conditions are statistically significant ($p < 0.001$). Conversely, those in the *flexible* conditions exercised significantly more outside of their workout windows than did participants in the *routine* conditions. Participants in the *flexible* conditions completed 0.98 out-of-window workouts per week on average and at least 1 out-of-window workout in a given week 48.5% of the time, whereas participants in the *routine* conditions completed 0.42 out-of-window workouts per week on average and at least 1 out-of-window workout in a given week 26.8% of the time. Tables 2, column (9) and 3, column (9) show that these comparisons are also statistically significant ($p < 0.001$).

Another way of analyzing the mix of in-window and out-of-window visits is to examine the fraction of participants' workouts that took place during their workout windows by experimental condition. For each participant, we calculate the number of weekdays during the incentive period that featured a gym visit during his or her workout window, and we divide that number by the number of weekdays during the incentive period that featured any gym visit. Dropping individuals for whom the denominator of the fraction is zero (i.e., individuals who did not visit the gym during the incentive period), we find that the mean of the fraction is 77.7% in the *routine* conditions. This is significantly higher than the 50.8% mean in the *flexible* conditions ($p < 0.001$) and the 53.2% mean in the *control* condition ($p < 0.001$).

3.3. Postintervention Results

3.3.1. Results for Levels of Exercise Activity. Patterns of postintervention gym attendance over our four-week follow-up period are depicted in Figure 3 in the online appendix. Specifically, the three plots in Figure 3A in the online appendix present means of overall, in-window, and out-of-window gym visits for each week by experimental condition, whereas the three plots in Figure 3B in the online appendix present the fraction of participants with at least one overall, in-window, and out-of-window gym visit for each week by experimental condition.[15] Tables 4 and 5 present the regression analogs of Figure 3 in the online appendix.

We replicate the well-established finding from Charness and Gneezy (2009), Acland and Levy (2015), and Royer et al. (2015) that when participants are paid to exercise repeatedly, they continue to exercise significantly more even after payments cease compared with a control group that is never paid to exercise. As Table 4, column (7) reports, participants in the *flexible* conditions made 0.25 more overall gym visits per week in the four-week postintervention period than participants in the *control* condition ($p < 0.05$). As Table 5, column (7) shows, a similar pattern emerges when we consider a participant's likelihood of working out at least once in a given week. Participants in the *flexible* conditions were 12 percentage points more likely to visit the gym at least once in a given week during our follow-up period than those in the *control* condition ($p < 0.001$).

We can also compare our *flexible* and *routine* conditions. Table 4, column (7) shows that participants in the *flexible* conditions had a marginally significant 0.10 more overall gym visits per week than participants in the *routine* conditions ($p < 0.10$), and Table 5, column (7) shows that participants in the *flexible* conditions had a 6-percentage point higher likelihood of visiting the gym at least once in a given week than participants in the *routine* conditions ($p < 0.001$). These differences in postintervention gym attendance are approximately twice the size of those induced by a \$4 increase in the incentive offered for qualifying gym visits during our intervention period (that is, the difference between our \$3 and \$7 incentive conditions), which are not statistically significant.

Although participants in the *routine* conditions worked out less frequently overall postintervention than those in the *flexible* conditions, they did not exhibit a statistically significantly different number of in-window gym visits (see Table 4, column (8)). Similarly, although participants in the *routine* conditions were less likely to make at least one gym visit in a given week postintervention than participants in the *flexible* conditions, Table 5, column (8) shows that participants in the *routine* and *flexible* conditions worked out at least once during their workout windows in a given week at similar rates.[16] To further explore this pattern, we count the number of weekdays during the four-week postintervention period on which a participant visited the gym during his or her workout window, and we divide by the number of weekdays on which the participant visited the gym at all. We call this the participant's fraction of in-window gym visits. Among participants who ever visited the gym during the four-week postintervention period, the mean fraction of in-window gym visits is highest in the *routine* conditions (55.3%). The mean fraction of in-window gym visits is significantly lower in the *flexible* conditions (47.0%). For participants in

Table 4. Regressions Predicting Participants' Weekly Workouts During Postintervention Weeks 1–4

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Total workouts	Total in-window workouts	Total out-of-window workouts	Total workouts	Total in-window workouts	Total out-of-window workouts	Total workouts	Total in-window workouts	Total out-of-window workouts
Flexible payment $3	0.21^{+}	0.07	0.13*						
	(0.11)	(0.08)	(0.07)						
Flexible payment $7	0.28*	0.09	0.20**						
	(0.11)	(0.08)	(0.07)						
Routine payment $3	0.12	0.07	0.04						
	(0.11)	(0.08)	(0.07)						
Routine payment $7	0.18	0.15^{+}	0.01						
	(0.11)	(0.08)	(0.07)						
$3 Interventions				0.17	0.07	0.09			
				(0.11)	(0.07)	(0.06)			
$7 Interventions				0.23*	0.12^{+}	0.10^{+}			
				(0.11)	(0.07)	(0.06)			
Flexible interventions							0.25*	0.08	0.16**
							(0.11)	(0.07)	(0.06)
Routine interventions							0.15	0.11	0.03
							(0.11)	(0.07)	(0.06)
Mean values of control group	0.76	0.42	0.39	0.76	0.42	0.39	0.76	0.42	0.39
Observations	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032
R^2	0.08	0.05	0.05	0.08	0.05	0.04	0.08	0.05	0.05
Wald test ($3 flexible – $7 flexible)									
Difference in coefficients	−0.07	−0.02	−0.06						
	(0.07)	(0.05)	(0.05)						
Wald test ($3 flexible – $3 routine)									
Difference in coefficients	0.09	0.00	0.09^{+}						
	(0.07)	(0.05)	(0.05)						
Wald test ($3 flexible – $7 routine)									
Difference in coefficients	0.04	−0.08	0.12**						
	(0.07)	(0.06)	(0.04)						
Wald test ($7 flexible – $3 routine)									
Difference in coefficients	0.16*	0.02	0.15**						
	(0.07)	(0.05)	(0.05)						

Table 4. (Continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Total workouts	Total in-window workouts	Total out-of-window workouts	Total workouts	Total in-window workouts	Total out-of-window workouts	Total workouts	Total in-window workouts	Total out-of-window workouts
Wald test ($7 flexible – $7 routine)									
Difference in coefficients	0.10 (0.07)	–0.06 (0.06)	0.18*** (0.05)						
Wald test ($3 routine – $7 routine)									
Difference in coefficients	–0.06 (0.07)	–0.09 (0.06)	0.03 (0.04)						
Wald test ($3 – $7)									
Difference in coefficients				–0.06 (0.05)	–0.05 (0.04)	–0.02 (0.03)			
Wald test (flexible – routine)									
Difference in coefficients							0.10^{+} (0.05)	–0.03 (0.04)	0.14*** (0.03)

Notes. This table reports a series of ordinary least squares regressions predicting a study participant's weekly number of (a) overall workouts, (b) workouts initiated during his or her workout window, and (c) workouts initiated outside of his or her workout window during the four weeks following the intervention period. In each column, we report the mean number of workouts completed by the control group within this period. The primary predictors included in these regressions are treatment status indicators, which indicate the size of the incentive offered for exercise ($3 vs. $7) and the flexibility of the workout schedule (flexible vs. routine). We report pairwise Wald tests to assess whether all paired regression coefficients reported differ significantly from each other. Standard errors clustered by workout buddy pair are in parentheses. The control variables in the regressions are indicators for randomization strata (12 strata: three randomization dates, crossed with whether workout window perfectly overlapped with partner's workout window, crossed with whether self-reported pair number of workouts per week was above median for randomization date), as well as an indicator for missing workout window.

$^{+}p<0.10$; $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$.

Table 5. Regressions Predicting Participants' Likelihood of Working Out Each Week During Postintervention Weeks 1–4

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Any workouts? (Y/N)	Any in-window workouts? (Y/N)	Any `out-of-window workouts? (Y/N)	Any workouts? (Y/N)	Any in-window workouts? (Y/N)	Any out-of-window workouts? (Y/N)	Any workouts? (Y/N)	Any in-window workouts? (Y/N)	Any out-of-window workouts? (Y/N)
Flexible payment $3	0.10**	0.04	0.07*						
	(0.04)	(0.03)	(0.03)						
Flexible payment $7	0.14***	0.07*	0.10**						
	(0.04)	(0.03)	(0.03)						
Routine payment $3	0.05	0.04	0.01						
	(0.04)	(0.03)	(0.03)						
Routine payment $7	0.07$^+$	0.06*	0.01						
	(0.04)	(0.03)	(0.03)						
$3 Interventions				0.08*	0.04	0.04			
				(0.03)	(0.03)	(0.03)			
$7 Interventions				0.10**	0.07*	0.05$^+$			
				(0.03)	(0.03)	(0.03)			
Flexible interventions							0.12***	0.05$^+$	0.08*
							(0.03)	(0.03)	(0.03)
Routine interventions							0.06$^+$	0.05$^+$	0.01
							(0.03)	(0.03)	(0.03)
Mean values of control group	0.34	0.22	0.23	0.34	0.22	0.23	0.34	0.22	0.23
Observations	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032
R^2	0.07	0.05	0.05	0.07	0.05	0.05	0.07	0.05	0.05
Wald test ($3 flexible – $7 flexible)									
Difference in coefficients	–0.04	–0.03	–0.03						
	(0.02)	(0.02)	(0.02)						
Wald test ($3 flexible – $3 routine)									
Difference in coefficients	0.05*	0.00	0.06**						
	(0.02)	(0.02)	(0.02)						
Wald test ($3 flexible – $7 routine)									
Difference in coefficients	0.03	–0.03	0.06**						
	(0.02)	(0.02)	(0.02)						

Table 5. (Continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Any workouts? (Y/N)	Any in-window workouts? (Y/N)	Any `out-of-window workouts? (Y/N)	Any workouts? (Y/N)	Any in-window workouts? (Y/N)	Any out-of-window workouts? (Y/N)	Any workouts? (Y/N)	Any in-window workouts? (Y/N)	Any out-of-window workouts? (Y/N)
Wald test ($7 flexible – $3 routine)									
Difference in coefficients	0.08*** (0.02)	0.03 (0.02)	0.08*** (0.02)						
Wald test ($7 flexible – $7 routine)									
Difference in coefficients	0.07** (0.02)	0.01 (0.02)	0.09*** (0.02)						
Wald test ($3 routine – $7 routine)									
Difference in coefficients	–0.02 (0.02)	–0.02 (0.02)	0.00 (0.02)						
Wald test ($3 – $7)									
Difference in coefficients				–0.03 (0.02)	–0.03$^+$ (0.02)	–0.01 (0.01)			
Wald test (flexible – routine)									
Difference in coefficients							0.06*** (0.02)	0.00 (0.02)	0.07*** (0.01)

Notes. This table reports a series of ordinary least squares regressions predicting a study participant's weekly likelihood of completing a (a) workout anytime, (b) workout initiated during his or her workout window, and (c) workout initiated outside of his or her workout window during the four weeks following the intervention period. In each column, we report the mean weekly fraction of participants in the control group who completed a workout within this period. The primary predictors included in these regressions are treatment status indicators, which indicate the size of the incentive offered for exercise ($3 vs. $7) and the flexibility of the workout schedule (flexible vs. routine). We report pairwise Wald tests to assess whether all paired regression coefficients reported differ significantly from each other. Standard errors clustered by workout buddy pair are in parentheses. The control variables in the regressions are indicators for randomization strata (12 strata: three randomization dates, crossed with whether workout window perfectly overlapped with partner's workout window, crossed with whether self-reported pair number of workouts per week was above median for randomization date), as well as an indicator for missing workout window. N, no; Y, yes.

$^+p<0.10$; $^*p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$.

the *control* condition, this statistic falls between the previous two at 51.3%. Overall, these patterns are consistent with the hypothesis that the *routine* conditions encouraged the formation of routines such that participants in these conditions developed a sustained habit of visiting the gym during their workout windows.

Participants in the *flexible* conditions completed 0.14 more out-of-window workouts per week postintervention than participants in the *routine* conditions ($p < 0.001$) and 0.16 more out-of-window workouts per week postintervention than participants in the *control* condition ($p < 0.01$). Similarly, participants in the *flexible* conditions were 7 percentage points more likely to make at least one out-of-window gym visit in a given week postintervention than participants in the *routine* conditions ($p < 0.001$) and 8 percentage points more likely to make at least one out-of-window gym visit in a given week postintervention than participants in the *control* condition ($p < 0.05$).

3.3.2. Results for Changes in Exercise Activity. To complement our analysis of levels of exercise activity during the four-week postintervention period, we also examine changes in exercise activity from the intervention period to the four-week postintervention period. In Table 6, the outcome variable is the change in a participant's mean weekly number of overall, in-window, or out-of-window gym visits from the intervention period to the four-week postintervention period. In Table 7, the outcome variable is the change in a participant's mean of the indicator for having at least one overall, in-window, or out-of-window gym visit in a given week. Relative to the $3 incentive conditions, the $7 incentive conditions exhibited larger decreases in exercise activity after incentives were removed. Of course, this result may not be surprising because the $7 incentive conditions exhibited more exercise activity during the intervention period (see Tables 2 and 3), implying that a return to baseline would represent a larger decrease.

It is more interesting to focus on a comparison of the *routine $7* condition and the *flexible $3* condition. These two conditions induced similar numbers of total workouts during the intervention period, but the workouts induced by these conditions during the intervention period were distributed differently: the *routine $7* condition produced more in-window workouts and fewer out-of-window workouts than the *flexible $3* condition. Past research on the psychology of habit formation suggests that the routine behavior induced by the *routine $7* condition should help to establish exercise habits and therefore, lead to less of a decrease in exercise activity after the end of the intervention period. However, Figure 2 shows that if anything, the opposite pattern emerged in the data, with the *routine $7* condition exhibiting a larger decrease in the number of overall workouts per week than the *flexible $3* condition. Table 6, column (1) indicates that this difference is statistically significant: the *routine $7* condition had a decrease in the number of overall workouts per week that was 0.14 larger than the one we observe in the *flexible $3* condition ($p < 0.05$). As Table 7, column (1) reports, the *routine $7* condition had a decrease in the likelihood of having at least one overall workout in a given week that was 5 percentage points larger than the decrease observed in the *flexible $3* condition ($p < 0.05$).

3.3.3. Results Using an Instrumental Variables Framework. We also implemented an instrumental variables strategy that predicts exercise activity in the postintervention period using exercise activity during the intervention. Specifically, the outcome variable in these regressions is the total number of gym visits per week during the four weeks immediately following the intervention period, or an indicator for having at least one gym visit during a given week, again limiting the sample to the four weeks immediately following the intervention period. We also conducted versions of the instrumental variables analysis that focus only on in-window visits or only on out-of-window visits during the four weeks following the intervention period. The right-hand-side variables of interest are the number of in-window gym visits and the number of out-of-window gym visits per week during the intervention period. We instrument for these two variables using four treatment group indicators, omitting an indicator for the control group. Table 19 in the online appendix shows the regression results. An incremental in-window gym visit per week during the intervention period leads to 0.22 extra total gym visits per week and a 9-percentage point increase in the likelihood of visiting the gym at least once in a given week during the four weeks following the intervention period. An incremental out-of-window gym visit per week during the intervention period leads to 0.32 extra total gym visits per week and a 16-percentage point increase in the likelihood of visiting the gym at least once in a given week during the four weeks following the intervention period. The difference between these coefficients is statistically significant when the outcome variable is the indicator for visiting the gym at least once in a given week. These results bolster our main finding that the *flexible* conditions, which generated more out-of-window and fewer in-window gym visits during the intervention when compared with the *routine* conditions, generated more exercise activity during the four weeks following the intervention period.

Table 6. Regressions Predicting the Change in Participants' Weekly Workouts from the Four-Week Intervention Period to Postintervention Weeks 1–4

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Change in total workouts	Change in total in-window workouts	Change in total out-of-window workouts	Change in total workouts	Change in total in-window workouts	Change in total out-of-window workouts	Change in total workouts	Change in total in-window workouts	Change in total out-of-window workouts
Flexible payment $3	−0.37***	−0.25***	−0.13						
	(0.10)	(0.07)	(0.08)						
Flexible payment $7	−0.61***	−0.34***	−0.31***						
	(0.10)	(0.07)	(0.08)						
Routine payment $3	−0.28**	−0.51***	0.19*						
	(0.10)	(0.07)	(0.08)						
Routine payment $7	−0.51***	−0.81***	0.22**						
	(0.10)	(0.08)	(0.08)						
$3 Interventions				−0.33***	−0.38***	0.03			
				(0.09)	(0.06)	(0.08)			
$7 Interventions				−0.56***	−0.57***	−0.05			
				(0.09)	(0.07)	(0.08)			
Flexible interventions							−0.49***	−0.29***	−0.22**
							(0.09)	(0.06)	(0.08)
Routine interventions							−0.40***	−0.66***	0.20**
							(0.09)	(0.07)	(0.08)
Mean values of control group	−0.35	−0.17	−0.21	−0.35	−0.17	−0.21	−0.35	−0.17	−0.21
Observations	2,508	2,508	2,508	2,508	2,508	2,508	2,508	2,508	2,508
R^2	0.03	0.07	0.08	0.03	0.04	0.01	0.02	0.06	0.08
Wald test ($3 flexible – $7 flexible)									
Difference in coefficients	0.24***	0.09^{+}	0.18***						
	(0.07)	(0.05)	(0.05)						
Wald test ($3 flexible – $3 routine)									
Difference in coefficients	−0.09	0.26***	−0.32***						
	(0.06)	(0.06)	(0.04)						
Wald test ($3 flexible – $7 routine)									
Difference in coefficients	0.14*	0.56***	−0.35***						
	(0.07)	(0.06)	(0.05)						

Table 6. (Continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Change in total workouts	Change in total in-window workouts	Change in total out-of-window workouts	Change in total workouts	Change in total in-window workouts	Change in total out-of-window workouts	Change in total workouts	Change in total in-window workouts	Change in total out-of-window workouts
Wald test ($7 flexible – $3 routine)									
Difference in coefficients	−0.33***	0.17**	−0.50***						
	(0.07)	(0.06)	(0.05)						
Wald test ($7 flexible – $7 routine)									
Difference in coefficients	−0.10	0.47***	−0.53***						
	(0.07)	(0.07)	(0.05)						
Wald test ($3 routine – $7 routine)									
Difference in coefficients	0.23***	0.30***	−0.03						
	(0.07)	(0.07)	(0.04)						
Wald test ($3 – $7)									
Difference in coefficients				0.24***	0.20***	0.07*			
				(0.05)	(0.04)	(0.03)			
Wald test (flexible – routine)									
Difference in coefficients							−0.10*	0.36***	−0.42***
							(0.05)	(0.04)	(0.03)

Notes. This table reports a series of ordinary least squares regressions predicting a study participant's change from the four-week intervention period to the four weeks following the intervention period in average weekly number of (a) overall workouts, (b) workouts initiated during his or her workout window, and (c) workouts initiated outside of his or her workout window. In each column, we report the mean change for the control group. The primary predictors included in these regressions are treatment status indicators, which indicate the size of the incentive offered for exercise ($3 vs. $7) and the flexibility of the workout schedule (flexible vs. routine). We report pairwise Wald tests to assess whether all paired regression coefficients reported differ significantly from each other. Standard errors clustered by workout buddy pair are in parentheses. The control variables in the regressions are indicators for randomization strata (12 strata: three randomization dates, crossed with whether workout window perfectly overlapped with partner's workout window, crossed with whether self-reported pair number of workouts per week was above median for randomization date), as well as an indicator for missing workout window.

$^{+}p<0.10$; $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$.

Table 7. Regressions Predicting the Change in Participants' Likelihood of Working Out Each Week from the Four-Week Intervention Period to Postintervention Weeks 1–4

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Change in mean of any workouts indicator	Change in mean of any in-window workouts indicator	Change in mean of any out-of-window workouts indicator	Change in mean of any workouts indicator	Change in mean of any in-window workouts indicator	Change in mean of any out-of-window workouts indicator	Change in mean of any workouts indicator	Change in mean of any in-window workouts indicator	Change in mean of any out-of-window workouts indicator
Flexible payment $3	−0.06^{+} (0.03)	−0.10*** (0.02)	−0.06^{+} (0.03)						
Flexible payment $7	−0.09** (0.03)	−0.10*** (0.02)	−0.10** (0.03)						
Routine payment $3	−0.08* (0.03)	−0.16*** (0.03)	0.05^{+} (0.03)						
Routine payment $7	−0.10*** (0.03)	−0.24*** (0.03)	0.08** (0.03)						
$3 Interventions				−0.07* (0.03)	−0.13*** (0.02)	−0.00 (0.03)			
$7 Interventions				−0.10*** (0.03)	−0.17*** (0.02)	−0.01 (0.03)			
Flexible interventions							−0.07** (0.03)	−0.10*** (0.02)	−0.08** (0.03)
Routine interventions							−0.09** (0.03)	−0.20*** (0.02)	−0.07* (0.03)
Mean values of control group	−0.16	−0.09	−0.09	−0.16	−0.09	−0.09	−0.16	−0.09	−0.09
Observations	2,508	2,508	2,508	2,508	2,508	2,508	2,508	2,508	2,508
R^2	0.02	0.05	0.05	0.02	0.03	0.00	0.02	0.04	0.05
Wald test ($3 flexible – $7 flexible)									
Difference in coefficients	0.03 (0.02)	0.01 (0.02)	0.04* (0.02)						
Wald test ($3 flexible – $3 routine)									
Difference in coefficients	0.02 (0.02)	0.07** (0.02)	−0.11*** (0.02)						
Wald test ($3 flexible – $7 routine)									
Difference in coefficients	0.05* (0.02)	0.14*** (0.02)	−0.14*** (0.02)						

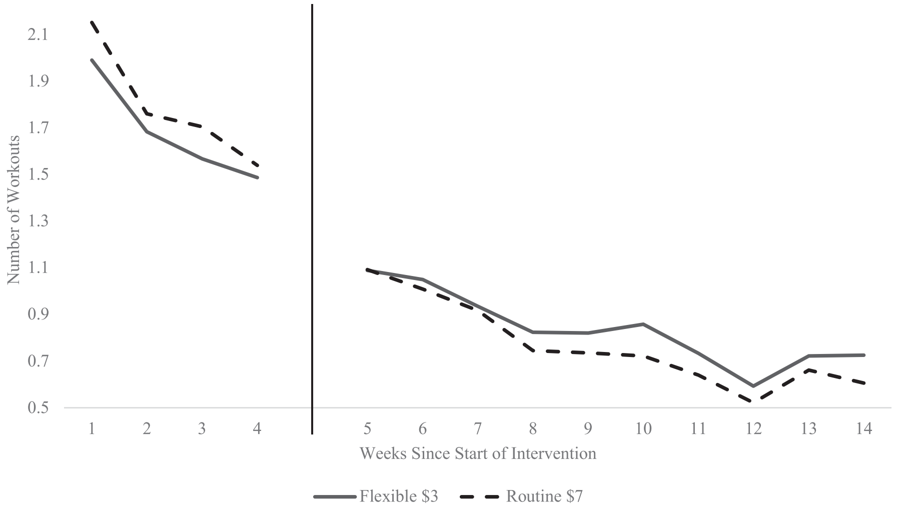
Table 7. (Continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Change in mean of any workouts indicator	Change in mean of any in-window workouts indicator	Change in mean of any out-of-window workouts indicator	Change in mean of any workouts indicator	Change in mean of any in-window workouts indicator	Change in mean of any out-of-window workouts indicator	Change in mean of any workouts indicator	Change in mean of any in-window workouts indicator	Change in mean of any out-of-window workouts indicator
Wald test ($7 flexible – $3 routine)									
Difference in coefficients	–0.01 (0.02)	0.06** (0.02)	–0.15*** (0.02)						
Wald test ($7 flexible – $7 routine)									
Difference in coefficients	0.01 (0.02)	0.13*** (0.02)	–0.19*** (0.02)						
Wald test ($3 routine – $7 routine)									
Difference in coefficients	0.03 (0.02)	0.07** (0.02)	–0.03$^+$ (0.02)						
Wald test ($3 – $7)									
Difference in coefficients				0.03* (0.02)	0.04* (0.02)	0.01 (0.01)			
Wald test (flexible – routine)									
Difference in coefficients							0.02 (0.02)	0.10*** (0.01)	–0.15*** (0.01)

Notes. This table reports a series of ordinary least squares regressions predicting a study participant's change from the four-week intervention period to the four weeks following the intervention period in average weekly likelihood of completing a (a) workout anytime, (b) workout initiated during his or her workout window, and (c) workout initiated outside of his or her workout window. In each column, we report the mean change for the control group. The primary predictors included in these regressions are treatment status indicators, which indicate the size of the incentive offered for exercise ($3 vs. $7) and the flexibility of the workout schedule (flexible vs. routine). We report pairwise Wald tests to assess whether all paired regression coefficients reported differ significantly from each other. Standard errors clustered by workout buddy pair are in parentheses. The control variables in the regressions are indicators for randomization strata (12 strata: three randomization dates, crossed with whether workout window perfectly overlapped with partner's workout window, crossed with whether self-reported pair number of workouts per week was above median for randomization date), as well as an indicator for missing workout window.

$^+p<0.10$; $^*p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$.

Figure 2. Comparing Exercise in the *Routine $7* and *Flexible $3* Conditions During the Intervention and Postintervention



Notes. This graph focuses on two experimental conditions that induced similar numbers of overall workouts during the intervention period but with different distributions of in-window vs. out-of-window workouts. The *routine $7* condition produced more in-window workouts and fewer out-of-window workouts than the *flexible $3* condition.

3.4. Heterogeneity Analyses

We conducted a number of heterogeneity analyses to determine whether our results in Tables 2–7 varied as a function of individual characteristics. We did not find statistically significant heterogeneity as a function of the time of day when a participant scheduled her workout window, whether a participant had the same workout window as her partner, or whether a participant was an above- versus below-median exerciser preintervention (based on the self-reported typical number of preintervention workouts completed weekly). We also did not find heterogeneity as a function of a participant's gender, BMI (based on self-reported weight and height), or job function (which might be related to the degree of flexibility in a participant's work schedule). We did find some statistically significant treatment effect heterogeneity by participant job level. However, further inspection revealed that this heterogeneity was driven by thinly populated job levels, and we therefore suspect that the finding is attributable to multiple hypothesis testing.

In addition to searching for heterogeneous treatment effects (statistically significant differences in treatment effects across subgroups), we searched for subgroups for which there was even suggestive evidence that the *routine* conditions led to more gym visits during the four-week postintervention period than the *flexible* conditions. Such a difference within a subgroup would be in the opposite direction of our results for the full sample, and our model of habit formation indicates that such subgroups might exist. However, when we examined subgroups defined by the individual characteristics studied in the previous paragraph (e.g., the time of day when a participant scheduled her workout window), only one subgroup—participants whose job function was "general and administrative"—offered evidence suggesting that the *routine* conditions led to more gym visits than the *flexible* conditions. In this case, the difference was nearly zero and was not statistically significant. Thus, we found essentially no evidence for situations in which routine incentives were more impactful than flexible incentives, but we cannot rule out the possibility that such situations exist.

3.5. Longevity of Effects

We also examined the longevity of the treatment effects by repeating our analyses using data on participants' gym visits during postintervention weeks 5–10 and postintervention weeks 11–40 (Figures 4 and 5 and Tables 2–5 in the online appendix). In addition, we tested the effect of having any incentives for exercise on the number of total gym visits per week and the likelihood of working out in a given week during postintervention weeks 5–10 and postintervention weeks 11–40 by regressing the total weekly workouts outcome variable and the indicator for having at least one workout in a given week on an indicator that takes the value of one if a participant was randomly assigned to one of the incentive conditions and the value of zero if a participant was in the control group (Table 6 in the online appendix).

As shown in Table 2 in the online appendix, during postintervention weeks 5–10, participants in the *flexible* conditions visited the gym 0.11 more times per week ($p < 0.01$) and had a 6-percentage point higher likelihood of a gym visit in a given week ($p < 0.001$)

than participants in the *routine* conditions. However, as shown in Table 4 in the online appendix, the differences in coefficients are no longer statistically significant during postintervention weeks 11–40. Table 6 in the online appendix shows that participants in the four incentive conditions pooled, relative to participants in the *control* condition, were 7 percentage points more likely to make a gym visit in a given week during postintervention weeks 5–10 ($p < 0.05$). They were also a marginally statistically significant 5 percentage points more likely to make a gym visit in a given week during postintervention weeks 11–40 ($p < 0.10$). Figure 6 in the online appendix displays the mean number of overall gym visits during each week for the four incentive conditions pooled and for the *control* condition. The difference was not statistically significant for this outcome measure when pooling postintervention weeks 5–10 (Table 6, column (1) in the online appendix) or when pooling postintervention weeks 11–40 (Table 6, column (3) in the online appendix).

3.6. Robustness Checks

We conducted a number of robustness checks. We found that our key results were qualitatively similar when we (a) introduced additional control variables into our regressions, (b) limited the sample to Google's largest office location, (c) only counted gym visits if a participant badged out at least 30 minutes after badging in, and (d) examined outcome variables capturing the minutes a participant spent at the gym per week. We also found that our key results were qualitatively similar when we limited the sample to participants who chose workout windows during the typical workday (that is, starting at 9:00 a.m. or later and ending at 5:00 p.m. or earlier). One possible concern with our main results is that participants in the *routine* conditions may only *appear* to have exercised less frequently during the four-week postintervention period than participants in the *flexible* conditions because participants in the *routine* conditions developed the habit of exercising during their workout windows and then sustained the habit by exercising *at home* during their workout windows, leading to fewer observed visits to workplace gyms. Participants in the *routine* conditions who chose workout windows during the typical workday probably did not exercise at home during their workout windows, so the finding that the results were similar when we limited the sample to participants with workout windows during the workday suggests that this alternative explanation does not drive our main results. Finally, we examined self-reported data on exercise outside of Google gyms. We did not find evidence that patterns in exercise outside of Google gyms offset experimental treatment effects on the frequency of visits to Google gyms during the intervention period or on the frequency of visits to Google gyms during the four weeks following the intervention period. However, when we compare the *routine $7* condition and the *flexible $3* condition, we cannot rule out the possibility that the difference in the change in Google gym visits from the intervention period to the four weeks following the intervention period is offset by the difference in the change in exercise outside of Google gyms. For further details, see Online Appendix F.

4. Discussion

4.1. A Model of the Trade-off Between Flexibility and Routinization

We present a simple model of habit formation that can help explain the patterns in our experimental data while also offering insight into the conditions under which flexible incentives may be more versus less effective than routine incentives at promoting habits.

4.1.1. Model Setup. In the model, there is one agent who makes decisions in two periods. In each period, the agent faces two opportunities to take an action, the in-window opportunity and the out-of-window opportunity.[17] For concreteness, the action might be visiting the gym, flossing one's teeth, or giving feedback to employees. The agent can take the action at most once per period, so there are three possible decisions a_t during period t: taking an in-window action ($a_t = I$), an out-of-window action ($a_t = O$), or no action ($a_t = N$). Period 1 is the intervention period, during which financial incentives for taking the action might be offered. Period 2 is the postintervention period, during which financial incentives do not apply. An in-window or out-of-window action in period 1 forms a habit, increasing the utility of that same type of action in period 2.

We normalize the utility of not taking the action (N) to zero in both periods. Relative to not taking the action, taking an in-window action and taking an out-of-window action are each associated with an intrinsic utility, defined as the net money-metric utility benefit that the agent receives from taking the action at that time, without accounting for any benefit from receiving financial incentives. We can think of the intrinsic utility as representing the personal enjoyment that the agent derives from taking the action at that time minus the opportunity cost of not engaging in some other activity at that time, but the intrinsic utility can, of course, capture many additional factors.

The intrinsic utility of an in-window action and the intrinsic utility of an out-of-window action in a given period are random variables, and they are drawn from known distributions and observed by the agent at the beginning of that period. The intrinsic utility of an in-window action in period 1 is $v_{in,1}$, the intrinsic utility of an out-of-window action in period 1 is $v_{out,1}$, the intrinsic utility of an in-window action in period 2 is $v_{in,2} + h_{in}(a_1)$, and the intrinsic utility of an out-of-window action in period 2 is $v_{out,2} + h_{out}(a_1)$. We assume that $v_{in,1}$, $v_{out,1}$, $v_{in,2}$, and $v_{out,2}$ are independent random variables, with $v_{in,1}$ and $v_{in,2}$ drawn from the uniform distribution on $\left[m_{in} - \frac{1}{2}, m_{in} + \frac{1}{2}\right]$ and with $v_{out,1}$ and $v_{out,2}$ drawn from the uniform distribution on $\left[m_{out} - \frac{1}{2}, m_{out} + \frac{1}{2}\right]$. We impose the restriction m_{in}, $m_{out} \in \left(-\frac{1}{2}, \frac{1}{2}\right)$ to ensure that the supports of the distributions always contain both strictly positive and strictly negative values. The terms $h_{in}(a_1)$ and $h_{out}(a_1)$ represent habit formation. We assume that taking an action of a certain type (in window versus out of window) in period 1 increases the intrinsic utility of taking an action of that same type in period 2 by h (i.e., $h_{in}(I) = h_{out}(O) = h$ and $h_{in}(O) = h_{in}(N) = h_{out}(I) = h_{out}(N) = 0$).[18] We impose the restriction $0 < h < min\{\frac{1}{2} - m_{in}, \frac{1}{2} - m_{out}\}$, again to ensure that the supports of the intrinsic utility distributions for in-window and out-of-window actions contain strictly negative values.

We compare the flexible incentive scheme and the routine incentive scheme with each other and with the control condition. Let i_{in} denote the incentive payment that the agent receives for an in-window action in period 1, i_{out} denote the incentive payment that the agent receives for an out-of-window action in period 1, and i_{no} denote the incentive payment that the agent receives for not taking the action in period 1. The flexible incentive scheme offers the agent $i_{in} = i_{out} = i_f$ and $i_{no} = 0$. The routine incentive scheme offers the agent $i_{in} = i_r$ and $i_{out} = i_{no} = 0$. In the control condition, we have $i_{in} = i_{out} = i_{no} = 0$. We impose the restriction $i_f, i_r \in \left(0, min\{\frac{1}{2} - m_{in}, \frac{1}{2} - m_{out}\}\right)$ to ensure that the intrinsic utility of an in-window or out-of-window action in period 1 plus the incentive that may be associated with such an action is sometimes strictly negative.

We assume that the agent is myopic in the sense that when he or she chooses an action in period 1, he or she does not consider the effect of habit formation on his or her expected utility in period 2.[19] Thus, in period 1, the agent compares $v_{in,1} + i_{in}$, $v_{out,1} + i_{out}$, and zero, and he or she chooses the option corresponding to the greatest among these (I, O, or N, respectively). In period 2, the agent compares $v_{in,2} + h_{in}(a_1)$, $v_{out,2} + h_{out}(a_1)$, and zero, and he or she chooses the option corresponding to the greatest among these (I, O, or N, respectively). Online Appendix G provides a complete characterization of this model. Here, we discuss the key conclusions from the model.

4.1.2. Predictions for Period 1. We first consider the model's predictions regarding the effect of the incentive schemes on decisions during the intervention (period 1), holding the dollar amount of the incentive offers constant across the flexible and routine schemes ($i_f = i_r = i$). Relative to the control condition, the flexible incentive scheme increases both the likelihood of an in-window action and the likelihood of an out-of-window action. The routine incentive scheme increases the likelihood of an in-window action by more than the flexible incentive scheme does, but it decreases the likelihood of an out-of-window action. On net, the flexible incentive scheme increases the likelihood of taking any action (in window or out of window) by more than the routine incentive scheme does. Intuitively, the routine incentive scheme promotes the in-window action both (a) in certain cases where the agent, in the absence of incentives, would have chosen to take neither the in-window action nor the out-of-window action and (b) in certain cases where the agent, in the absence of incentives, would have chosen the out-of-window action. The former effect represents an increase in the likelihood of taking any action (in window or out of window), but the latter effect merely represents a shift from an out-of-window action to an in-window action. The flexible incentive scheme, on the other hand, promotes the in-window action in certain cases where the agent, in the absence of incentives, would have chosen to take neither the in-window action nor the out-of-window action, and it also promotes the out-of-window action in certain cases where the agent, in the absence of incentives, would have chosen to take neither the in-window action nor the out-of-window action. The former effect corresponds to the first effect (a) of the routine incentive scheme, but the latter effect also represents an increase in the likelihood of taking any action (in window or out of window), thereby accounting for the greater impact of the flexible incentive scheme relative to the routine incentive scheme on the likelihood of taking any action. These predictions from the model are borne out in the experimental data.

4.1.3. Predictions for Period 2, Holding Incentive Dollar Amounts Constant. We now turn to the model's predictions regarding the effect of the incentive schemes on decisions postintervention (period 2), again holding the dollar amount of the incentive offers constant across the flexible and routine schemes ($i_f = i_r = i$). We focus on the likelihood of taking any action (in window or out of window) as the outcome of interest because this outcome reveals the relevant trade-offs associated with using flexible versus routine incentive schemes to create habits.

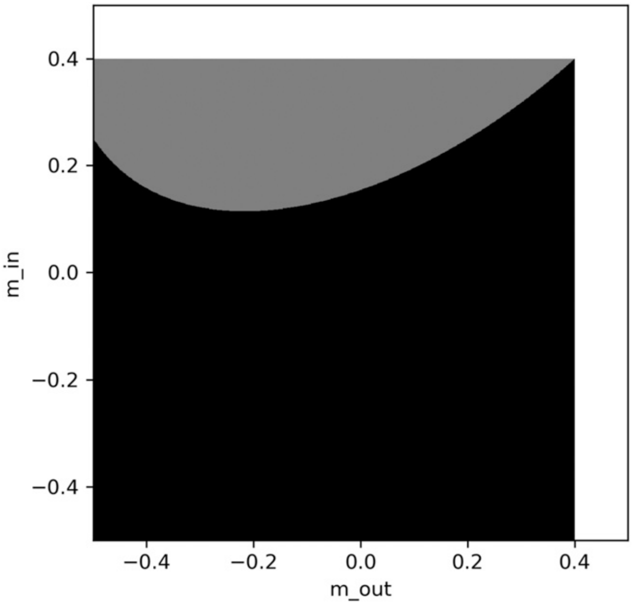
It is ambiguous whether the flexible incentive scheme or the routine incentive scheme will cause a larger increase in the likelihood of taking any action.[20] The sign of the comparison depends on the values for the parameters m_{in}, m_{out}, and i, but not h. In

Figure 3, we fix the value for the parameter i at 0.1 and display all combinations of values for m_{in} and m_{out} that satisfy our parameter restrictions. The combinations of m_{in} and m_{out} for which the flexible incentive scheme causes a larger increase in the likelihood of taking any action in period 2 than the routine incentive scheme are shaded in black, and the combinations for which the opposite is true are shaded in grey. The white areas denote combinations that do not satisfy our parameter restrictions.[21]

Figure 3 shows that the flexible incentive scheme causes a larger increase in the likelihood of taking any action in period 2 than the routine incentive scheme if m_{out} is greater than m_{in}—the entire area below the 45° line is shaded black. However, $m_{in} > m_{out}$ is not sufficient for the opposite to be true. For most possible values of m_{out}, m_{in} must be substantially higher than m_{out} in order for the routine incentive scheme to cause a larger increase in the likelihood of taking any action.

To develop intuition for this pattern, first consider how an in-window action or an out-of-window action in period 1 changes the likelihood of taking any action in period 2. An in-window action in period 1 increases the likelihood of an in-window action in period 2 because of habit formation. If m_{out} is low, the incremental in-window action in period 2 is likely to occur when the agent would not otherwise have taken any action, but if m_{out} is high, the incremental in-window action in period 2 is likely to replace an out-of-window action that would otherwise have occurred. The former effect represents an increase in the likelihood of taking any action, whereas the latter does not. Symmetric statements hold for the effect of an out-of-window action in period 1 on the likelihood of an out-of-window action in period 2. An incremental in-window action in period 1 therefore increases the likelihood of taking any action in period 2 more than an incremental out-of-window action in period 1 if and only if $m_{in} > m_{out}$.

Figure 3. The Relative Effect of Routine and Flexible Incentives in Our Simple Model of Habit Formation



Notes. The value of m_{in} varies along the vertical axis, and the value of m_{out} varies along the horizontal axis. The value of i is fixed at 0.1. Routine incentives increase the likelihood of taking any action (in window or out of window) postintervention by more than flexible incentives in the grey region, whereas flexible incentives have a greater effect in the black region. In the white region, $i > \min\{\frac{1}{2} - m_{in}, \frac{1}{2} - m_{out}\}$, so the incentive size is too large to be valid.

Now consider the effect of the incentive schemes on the likelihood of an in-window or an out-of-window action in period 1. The flexible incentive scheme increases both the likelihood of an in-window action and the likelihood of an out-of-window action. The routine incentive scheme, on the other hand, increases the likelihood of an in-window action by more than the flexible incentive scheme does while decreasing the likelihood of an out-of-window action. Thus, for the routine incentive scheme to cause a larger increase in the likelihood of taking any action in period 2 than the flexible incentive scheme, it must be the case that the effect of a period 1 in-window action on the likelihood of taking any action in period 2 is much greater than the effect of a period 1 out-of-window action. Based on the argument in the previous paragraph, satisfying this condition requires that m_{in} be substantially greater than m_{out}. See Online Appendix G for further details.

In the experiment, the flexible scheme generated more postintervention exercise than the routine scheme with the same dollar amount of incentives offered. Mapping the data to the model, the experimental setting was not a case in which m_{in} was substantially greater than m_{out}.

4.1.4. Predictions for Period 2, Holding Period 1 Activity Constant. Instead of holding the dollar amount of the incentive offers constant across the flexible and routine schemes, we can hold constant the likelihood of taking any action in period 1 and compare the likelihood of taking any action in period 2 for the flexible and routine incentive schemes. This comparison requires the dollar amount of the routine incentive offer to be larger than the dollar amount of the flexible incentive offer. In this case, the likelihood of taking any action in period 2 is greater for the flexible incentive scheme than for the routine incentive scheme if and only if $m_{in} < m_{out}$. Intuitively, relative to the flexible incentive scheme, the routine incentive scheme simply increases the likelihood of an in-window action in period 1 while decreasing the likelihood of an out-of-window action in period 1 by the same amount. The effect of the routine incentive scheme relative to the flexible incentive scheme on the likelihood of taking any action in period 2 therefore hinges on whether a period 1 in-window action or a period 1 out-of-window action exerts a stronger influence on the likelihood of taking any action in period 2. By the argument given in Section 4.1.3, this comparison is driven by the relative sizes of m_{in} and m_{out}.

In the experiment, the gym visit frequencies in the *flexible $3* and *routine $7* experimental conditions were approximately equal during the intervention. Tables 6 and 7 and Figure 2 indicate that, if anything, the *routine $7* condition exhibited a larger decrease in gym visit frequency postintervention than the *flexible $3* condition. Mapping this finding to the model, the experimental results suggest that m_{in} is slightly less than m_{out}.

4.1.5. Interpretation and Implications of the Model. The model offers guidance as to the types of activities for which flexible incentives versus routine incentives might be more effective for promoting habits. For some activities, there are regularly occurring opportunities for taking action that are frequently the most convenient or the most rewarding. For example, many individuals find it convenient to floss their teeth right before going to sleep for the night. When we apply the model to such activities, we would use parameters such that $m_{in} > m_{out}$, and we would predict that routine incentives would be more effective than flexible incentives for promoting habits. For other activities, the best opportunity for taking action occurs on an irregular schedule and under inconsistent circumstances. For example, the best opportunity for a manager to give developmental feedback to an employee may be when there is a temporary drop in the team's workload, but such drops may be the result of unpredictable decreases in the number of requests from clients. When we apply the model to these types of activities, we would use parameters such that $m_{in} < m_{out}$, and we would predict that flexible incentives would be more effective than routine incentives for promoting habits.

If we focus on a specific activity for which habit formation is desirable, the model also offers guidance as to the types of decision-making environments in which flexible incentives versus routine incentives might be more effective for promoting habits. Consider the case of promoting exercise habits in our experiment. Experimental participants were required to select a daily two-hour window to be the in-window gym visit opportunity, so the intrinsic utility of the in-window action is interpreted as the intrinsic utility of visiting the gym during that window, whereas the intrinsic utility of the out-of-window action is interpreted as the intrinsic utility of visiting the gym at whichever time outside that window is most desirable. It is plausible to anticipate $m_{in} > m_{out}$ in some environments and to anticipate $m_{in} < m_{out}$ in other environments. If an individual's day-to-day schedule is predictable and stable, there may be a two-hour window that is very frequently the best time to visit the gym, suggesting that $m_{in} > m_{out}$. On the other hand, if an individual's schedule varies significantly from one day to the next, the two-hour window that is most frequently the best time to visit the gym may still quite regularly be inferior to another time on a given day (it is simply a different time each day that is superior), suggesting that $m_{in} < m_{out}$. This latter description seems to apply to the participants in our experiment, whose workplace environment is dynamic and fast paced. In the model, the routine incentive scheme causes a larger increase in the likelihood of taking any action in period 2 than the flexible incentive scheme only when m_{in} is substantially greater than m_{out}, so the model implies that the routine scheme is better for habit formation than the flexible scheme when an individual's schedule is predictable and stable. This implication is important for managers and policy makers. Incentives for routines may be most impactful when they are applied in predictable and stable environments or when they are accompanied by a restructuring of the environment that creates opportune moments for taking action on a regular basis.

It would be possible to extend the model in several ways to illuminate other factors that may influence the comparison between flexible and routine incentive schemes. First, the interpretation of the in-window opportunity as a small window of time and the out-of-window opportunity as the most desirable among many alternative windows of time suggests that the parameter governing habit strength should be higher for the in-window action than for the out-of-window action (i.e., $h_{in}(I) = \bar{h} > h_{out}(O) = \underline{h}$). Such an assumption would reflect past research findings that successful habits are built on stable cues. This assumption would increase the effect of the routine incentive scheme on the likelihood of taking any action in period 2, but provided that $\bar{h}$ is not too much greater than $\underline{h}$, the model's implications based on the relative sizes of m_{in} and m_{out} would not change.

Second, the model could be extended by endogenizing the decision of which opportunity to label the in-window opportunity. If an individual is uncertain as to which opportunity is most frequently the best for taking the action, the flexible incentive scheme encourages more exploration than the routine incentive scheme and may therefore be more likely to help the individual discover a regular time that is particularly good for taking the action. Such a discovery may lead to more consistent postintervention engagement in the desired behavior. See Larcom et al. (2017) for evidence that forcing individuals to experiment with different routines can lead them to switch to more beneficial routines.[22]

Third, it would be natural to extend the model to endogenize the length of the time window associated with the in-window opportunity. An interesting trade-off arises in this extension. On one hand, increasing the length of the time window associated with the in-window opportunity increases m_{in} because a longer time window creates more opportunities for a high realization of the intrinsic utility of the in-window action. An increase in m_{in} increases the effectiveness of the routine incentive scheme relative to the flexible incentive scheme. On the other hand, if we allow the parameter governing habit strength to be higher for the in-window action than for the out-of-window action (i.e., $h_{in}(I) = \bar{h} > h_{out}(O) = \underline{h}$), increasing the length of the time window associated with the in-window opportunity is likely to decrease $\bar{h}$ because the in-window action becomes less strongly connected to a narrowly defined routine. A decrease in $\bar{h}$ decreases the effectiveness of the routine incentive scheme relative to the flexible incentive scheme.

Finally, it would be interesting to extend the model to consider different types of habit formation. For example, in the context of our experiment, if an individual is unable to visit the gym during the preselected workout window because of a scheduling conflict, having the commitment to figure out another time to go to the gym may be a habit-forming activity. The flexible incentive scheme encourages this behavior by rewarding out-of-window gym visits, so the flexible scheme may have the advantage that it promotes the resilience to find an alternative time to exercise in the face of scheduling conflicts.

Although our experiment is not designed to disentangle the exact mechanisms by which the flexible and routine incentive schemes exert influence on postintervention exercise, the results offer an important lesson for managers and policy makers who wish to help individuals form beneficial habits. Despite research indicating that successful habits are often characterized by engagement in a behavior under routine conditions, interventions designed to take advantage of this pattern face countervailing forces that may render them ineffective. The model provides insight into the types of activities and decision-making environments for which flexible versus routine incentive schemes are likely to be more impactful.

4.2. Limitations

In spite of its scale and scope, our study has a number of important limitations. First, we cannot perfectly measure participant exercise. In particular, we did not directly observe participants' exercise habits outside of Google's gyms, and some visits to Google gyms were unobserved because participants failed to badge in. We asked participants about these issues in our exit survey (see Section 3.6) and found that their responses generally did not undermine our main conclusions, but the self-reported information may not be reliable. Furthermore, a potential concern is that participants might collude with their workout partners or others in order to game the incentive system: for example, by bringing another employee's identification badge to the gym and recording a gym visit for that employee even when that employee did not visit the gym. Such behavior is unlikely to have occurred, however, because employees use their identification badges many times during the day and must keep them on hand in order to access each of the many different physical spaces within a Google campus.[23]

Second, our empirical results would likely have been different if we had made different decisions regarding the details of our experimental design. For example, all participants in the experiment, including those in the *control* condition, were asked to select a two-hour workout window that applied to every weekday. The routine incentive schemes may have been more effective if the windows were longer or shorter, if the windows were allowed to vary across days, if the windows could be adjusted according to work schedules or exercise class schedules,[24] or if the windows could be adjusted after participants had a chance to learn about the exercise times that worked best for them. Our experiment also did not attempt to "piggyback" exercise habits on top of existing routines, which may have been more effective (Judah et al. 2013), although it is not clear that existing routines could have been practicably harnessed for this purpose. Instead of using "piggybacking," our experiment involved sending an email reminder associated with each of the workout windows to cue exercise behavior. When many of these reminders failed to trigger a gym visit, participants may have felt discouraged, undermining the habit-forming potential of the intervention (although notably, our treatment conditions did produce lasting behavior change relative to our control condition).[25]

Furthermore, the intervention period only lasted for four weeks. Although this length of time matches the intervention duration in several previous experiments studying exercise habits (Charness and Gneezy 2009, Acland and Levy 2015, Royer et al. 2015), a longer intervention may be necessary to establish in-window exercise routines. In addition, all participants in our experiment were paired with a workout partner, and the effects of the intervention might have differed had participants instead signed up alone. For instance, although there was no obligation to coordinate with the workout partner, participants may have nonetheless tried to coordinate, perhaps in ways that made gym visits less convenient and thereby, undermined the formation of in-window exercise routines.[26] All of these experimental details are practical design considerations that a manager or policy maker who is seeking to promote exercise habits among employees or other populations must confront, so it would be valuable for future research to explore adjustments to these design features.

A third limitation of our study is that it was conducted at a single company (Google) with an employee population that is not representative of the U.S. workforce. Google has workplace gyms, and our findings might have differed if we had conducted the study at a gym that was not located at participants' place of work. Although we found no evidence of heterogeneous treatment effects by job function, the impact of routine versus flexible incentives might be different in organizations that structure work more or less flexibly. In general, employees at Google have higher levels of education and higher incomes than the average U.S. worker. Perhaps routine incentives did not generate persistent exercise habits because these high achievers had already established exercise routines prior to participating in the experiment.

Finally, the finding that routine incentives generated weaker exercise habits than flexible incentives might have been driven by participants' inferences regarding informational signals sent by the employer that were embedded in the design of the intervention. The informed consent form for our study explained that the experiment was conducted by outside academic researchers and that individual-level data collected during the course of the experiment would not be shared with Google, but participants might also have exhibited experimenter demand effects based on their understanding of the researchers' desired outcomes. Whether because of perceptions of the employer's desired outcomes or because of perceptions of the researchers' desired outcomes, perhaps participants in the *routine* conditions responded by visiting the gym during their workout windows even when doing so was highly inconvenient, undermining the success of habit formation. We view this possibility as a legitimate component of the *routine* conditions for judging their efficacy. After all, the *routine* conditions were intended to increase in-window gym visits in some situations where those gym visits would not have occurred in the absence of the intervention. The interpretation of the results is slightly different, but the results still speak to the likely effects of similar employer-sponsored programs to promote exercise routines because the introduction of any such programs may be accompanied by changes in perceptions regarding the employer's desired outcomes.[27]

5. Conclusion

In a large field experiment, we found that routine incentives, which offered monetary rewards for visiting the gym during a two-hour window, generated more gym visits during that window but fewer gym visits overall than flexible incentives, which offered monetary rewards for visiting the gym at any time. After the incentives were no longer offered, the participants who had received routine incentives exhibited less exercise activity than the participants who had received flexible incentives, consistent with past research showing that more repetition creates stronger habits. Our more important and novel contribution to the literature on habits focused on comparing participants who received large routine incentives ($7 per qualifying gym visit) and participants who received small flexible incentives ($3 per qualifying gym visit). These two groups visited the gym at a similar frequency during our intervention, but those in the routine group visited the gym at more consistent times. Comparing these two groups, we surprisingly find that participants who received large routine incentives subsequently exhibited larger postintervention decreases in exercise. Thus, despite past research suggesting that repeatedly rewarding beneficial behaviors under routine conditions might promote more lasting habits than repeatedly rewarding such behaviors on a flexible schedule, we find evidence for the other side of the trade-off: an incentive program that promotes rigid routines can be counterproductive to habit formation. Our simple model of habit formation suggests that routine incentives are unlikely to be more effective than flexible incentives in dynamic, fast-paced work environments but may be more successful in stable environments, where they can reinforce the development of routines that are less prone to disruption.

Our study raises a number of important questions for future research. We examined flexible incentive schemes and routine incentive schemes, but there may be a middle ground that is more effective than either of these options. For instance, an incentive scheme that pays participants for all workouts but pays *more* for in-window workouts might help individuals build an exercise routine while still encouraging participants who miss their workout window to exercise at another time. It would be valuable to explore this possibility further. In addition, we defined routines at a daily (rather than weekly or monthly) interval and defined workout windows as two-hour periods. Altering some of these definitions might have yielded different results. Finally, a routine incentive may be more or less effective in a social context than in an individual context. Workout partners who must stay on the same schedule to earn incentives may provide extra support and accountability to each other and make workouts more enjoyable, thereby making routines more persistent than they would otherwise be. On the other hand, a workout partner's failure to exercise may also license an individual to skip a scheduled gym visit, so a social routine could be less persistent. Future research should examine this issue and related questions to identify effective approaches for promoting long-term habit formation.

Acknowledgments

The authors thank Quoc Dang Hung Ho, Andrew Joung, David Mao, Predrag Pandiloski, Byron Perpetua, and Kartikeya Vira for outstanding research assistance. They also thank Google for integral support on the design and execution of this experiment. Finally, the authors are grateful for outstanding feedback from Susanna Gallani, Matthew Levy, Yan Chen (department editor), an associate editor, and two anonymous reviewers; from seminar audiences at Carnegie Mellon University, Columbia University, Cornell University, Harvard University, Northwestern University, Stanford University, University of Michigan, University of Pennsylvania, and Yale University; and from participants in the Behavioral Decision Research in Management Conference, the Advances in the Science of Habits Conference, the Center for Health Incentives and Behavioral Economics Conference, the Society for Judgment and Decision Making Conference, and the Behavior Change for Good Conference. This study is preregistered at ClinicalTrials.gov (NCT02346799) and is registered at the American Economic Association Randomized Controlled Trial Registry (AEARCTR-0003471). This research was approved by the institutional review boards of Harvard University, the National Bureau of Economic Research, and the University of Pennsylvania.

Endnotes

[1] For further references, see Thaler and Benartzi (2004), Milkman et al. (2014), Sen et al. (2014), Patel et al. (2016), and Staats et al. (2017).

[2] Indeed, in a survey of 69 psychology professors, 77% of respondents predicted that an individual who was induced to exercise at a regular time of day over the course of a month would form a more persistent habit than an individual who was induced to exercise the same amount over the course of a month but not necessarily at a regular time of day. However, this evidence is only suggestive because the survey asked about a hypothetical scenario that did not reflect all of the features of our field experiment. See Online Appendix A for details.

[3] We use the term "intrinsic utility" to denote utility from all sources other than financial incentive payments. We are not attempting to draw a connection to the distinction between intrinsic and extrinsic motivation.

[4] It may sound puzzling at first for the likelihood of an in-window action to be less than the likelihood of an out-of-window action. We interpret the opportunity for an in-window action as a narrow window of time (e.g., two hours during a day), whereas the opportunity for an out-of-window action encompasses several such windows of time (e.g., all other hours during the day). Thus, the in-window opportunity may be the preferred time to take the action more frequently than any other narrow window, but the likelihood of an in-window action may nonetheless be less than the likelihood of taking the action in any of several alternative windows.

[5] Employees were given a list of 96 two-hour time windows (one window starting every 15 minutes) and were told to select one. They were encouraged to discuss this time window with their work group to confirm that exercising during the time window would not be disruptive to their work.

[6] In our initial survey, 117 individuals had missing observations for their window selections. For the purposes of our stratified randomization procedure, we created a stratifying variable that was an indicator for perfect overlap with partner's window, coded as one if workout partners had perfectly matching (identical) selections of workout windows or if they both had a missing selection in the initial survey. We use this variable when constructing the strata fixed effects that serve as control variables in our regression analyses. However, we manually inputted workout windows for 116 of the 117 individuals after the randomization procedure. Workout windows including these updates are summarized in Tables 2 and 3 and are used to determine whether a given gym visit is an in-window or out-of-window visit. Our regressions control for an indicator for missing workout window, so the one individual with a missing window is effectively excluded from the analysis.

[7] At the request of our corporate partner, we also included four questions about overall well-being. Prior to the initiation of data collection, our research team committed to exclude these questions from our eventual analysis because they were not variables of interest to our team.

[8] We performed our power calculations using the online tool available at http://www.sample-size.net/means-effect-sizeclustered/. This calculator accounts for the effect of intracluster correlation on statistical power. Prior to collecting data, we assumed an intracluster correlation of 0.05 (a typical assumption) and a mean outcome of one gym visit per week, which gave us 80% power to detect a 33% difference in weekly exercise between the *control group* and the *flexible $3 payment group* and an 18.5% difference between treatment conditions. When we updated our power calculations postexperiment using the observed intracluster correlation in our sample of 0.26, the observed outcome standard deviation in the *control group* of 1.45, and the actual sample sizes, we determined that the detectable effect sizes in our study were 44% and 27%, respectively.

[9] Although participants were told that they would be required to finish both steps of the gym registration process (online and in-person registration) to be included in the study, randomization occurred as long as both partners had completed the online registration process. The rationale behind this decision was that upon first visiting the gym after online registration, participants would be automatically prompted to complete in-person registration, thus ensuring we would be able to track all gym visits. Of the 2,508 participants who were randomized to experimental conditions, 1,111 had not yet completed the in-person registration process by the date of their randomization (704 for the first randomization wave, 375 for the second, and 32 for the third). Participants who had not completed in-person registration received multiple reminder emails encouraging them to do so as soon as possible (see Online Appendix B, Figure B10).

[10] At the bottom of the daily reminder emails, participants were given links that would allow them to unsubscribe from the email and text message reminders.

[11] Note that to earn incentive payments for workouts, participants were required to badge out of the gym at least 30 minutes after badging in, so we use a more inclusive definition of a gym visit in our

analysis than in our rewards scheme. We believe that the inclusive definition of a gym visit better reflects an individual's exercise behavior. However, Tables 13–15 in the online appendix show that the results are similar if we use the less inclusive definition, which only counts a gym visit as having occurred if we see a study participant badge out of the gym at least 30 minutes after badging in.

[12] This occurs in 2.56% of our weekly observations. Although our decision to code variables in this way means that our statistical results regarding in-window and out-of-window gym visits do not "add up" to our statistical results regarding total gym visits, we believe that our variable definitions provide the best representation of the experimental results. When a participant has multiple employee identification badge swipes at the gym on the same day, the amount of time between swipes is less than two hours in the majority of cases, suggesting that two adjacent swipes are in fact associated with the same gym visit (perhaps with a break outside the gym in the middle of the visit). Thus, if a participant has both an in-window badge swipe and an out-of-window badge swipe on the same day, we record one gym visit when counting total gym visits, but because the gym visit straddles the exercise window boundary, we record one in-window gym visit when counting in-window gym visits and one out-of-window gym visit when counting out-of-window gym visits.

[13] The mean self-reported typical number of workouts per week is more than double the mean number of observed gym visits per week in the control group during the incentive period (we do not have data on badging in and badging out at the gym prior to the incentive period). Perhaps individuals have inflated perceptions of their own workout frequency or are reporting their ideal workout frequency. It is also possible that their responses incorporate workouts that do not take place at the gym.

[14] We conducted 10,000 simulations in which we randomly assigned individuals to pairs, holding fixed each individual's chosen workout window. Across the simulations, the mean fraction of pairs with exactly overlapping workout windows was 4%, and the range from the 2.5th percentile to the 97.5th percentile of the distribution of the fraction across simulations did not contain the observed fraction using the real pairings.

[15] It is interesting to note that the frequency of exercise declines from week to week during the postintervention period even in the control group. The experiment is not designed to explain this pattern, but perhaps the decline is because of a Hawthorne effect fading away.

[16] Participants in the *control* condition did not have a statistically significantly different mean number of postintervention in-window gym visits compared with participants in the *flexible* conditions or participants in the *routine* conditions. They were marginally significantly less likely to have at least one in-window gym visit in a given week postintervention ($p<0.10$).

[17] The decision of which opportunity to label the in-window opportunity and which opportunity to label the out-of-window opportunity is outside the model. When mapping the model to the experiment, we think of the in-window opportunity as the two-hour window that the agent expects to be the best window for visiting the gym, and we think of the out-of-window opportunity as the best opportunity among all other windows. After presenting the baseline version of the model, we discuss an extension that endogenizes the determination of the in-window and out-of-window opportunities.

[18] Instead of assuming that $h_{in}(I) = h_{out}(O) = h$, we could have assumed that $h_{in}(I) = \bar{h} > h_{out}(O) = \underline{h}$. It would also be possible to assume that taking an action of one type (in window or out of window) has a habit-forming effect on subsequently taking an action of the opposite type (out of window or in window, respectively; i.e., $h_{in}(O) = h_{out}(I) = h' > 0$). We decided not to pursue these approaches because they would add complexity to the model and would yield only incremental insights. If we were to make these alternative assumptions with $\bar{h}$ not too much greater than $\underline{h}$ and h' not too much greater than zero, we would draw qualitatively similar conclusions from the model.

[19] We have also analyzed the model with a sophisticated agent, who anticipates the impact of his or her period 1 action on his or her expected utility in period 2. The results are qualitatively similar.

[20] For all parameter values that we consider, the flexible incentive scheme causes an increase in the likelihood of taking any action in period 2, relative to the control group. For certain parameter values, the routine incentive scheme causes a decrease in the likelihood of taking any action. See Online Appendix G.

[21] In Online Appendix G, we recreate Figure 3 but also show analogous figures with the value for the parameter i changed to 0.05 or 0.2. Together, these three figures demonstrate that varying the value for the parameter i only slightly changes the comparison between the flexible and routine incentive schemes.

[22] To explore this possibility in the data from our experiment, we first identify the weekdays on which a given participant had an out-of-window gym visit during the four-week postintervention period. For each participant, we then calculate the fraction of those days that featured an out-of-window gym visit that could be matched to an intervention-period gym visit by the same participant meeting two criteria: (1) occurred on the same day of the week and (2) had the same starting time of day, plus or minus 15 minutes. Among participants who had at least one out-of-window gym visit during the four-week postintervention period, the mean of the fraction in the flexible conditions was 34.7%, which was larger than the 26.3% mean fraction in the routine conditions ($p<0.001$) but not statistically significantly different from the 31.6% mean fraction in the control condition. The differences across conditions are similar if we use a window of plus or minus 5 minutes or a window of plus or minus 30 minutes around the starting time of day.

[23] We also empirically examine whether this form of collusion might have occurred. For each participant, we calculate the fraction of intervention-period gym visits that started within five minutes of a gym visit by the workout partner. If collusion between workout partners was frequent, we would expect this fraction to be higher in the experimental conditions that make such collusion more financially beneficial. However, the mean of this fraction does not significantly vary across experimental conditions in an F test of joint equality.

[24] We do not have data on work schedules or exercise class schedules.

[25] Even among participants in the *routine $7* experimental condition, who had the most in-window gym visits during the intervention period, 69% of weekdays during the intervention were not associated with an in-window gym visit.

[26] To explore this possibility empirically, we separate each pair of participants into the member with more in-window gym visits and the member with fewer in-window gym visits during the intervention period. The individuals in the first category are less likely to have made inconvenient schedule adjustments to coordinate with their partners. However, when we conduct the analyses in Tables 2–7 using only this subset of the sample, the results are similar. Separating each participant pair based on the fraction of intervention-period gym visits that were in-window also delivers similar results. These patterns do not support the hypothesis that our main results are driven by participants' decisions to coordinate with their partners at the expense of convenience, but we do not rule out the hypothesis because the empirical tests are imperfect.

[27] Another concern is that experimenter demand effects might have been particularly strong during the first week of the intervention period, and data from that first week might be driving the results. However, when we conduct the same analysis as in Tables 2 and 3 but drop data from the first week of the intervention period, the results are similar.

References

Acland D, Levy MR (2015) Naiveté, projection bias, and habit formation in gym attendance. *Management Sci.* 61(1):146–160.

Becker GS, Murphy KM (1988) A theory of rational addiction. *J. Political Econom.* 96(4):675–700.

Benartzi S, Beshears J, Milkman KL, Sunstein CR, Thaler RH, Shankar M, Tucker W, Congdon WJ, Galing S (2017) Should governments invest more in nudging? *Psych. Sci.* 28(8):1041–1055.

Beshears J, Choi JJ, Laibson D, Madrian BC (2013) Simplification and saving. *J. Econom. Behav. Organ.* 95:130–145.

Brooks TL, Leventhal H, Wolf MS, O'Conor R, Morillo J, Martynenko M, Wisnivesky JP, Federman AD (2014) Strategies used by older adults with asthma for adherence to inhaled corticosteroids. *J. General Internal Medicine* 29(11): 1506–1512.

Carels RA, Young KM, Koball A, Gumble A, Darby LA, Oehlhof MW, Wott CB, Hinman N (2011) Transforming your life: An environmental modification approach to weight loss. *J. Health Psych.* 16(3):430–438.

Carels RA, Burmeister JM, Koball AM, Oehlhof MW, Hinman N, LeRoy M, Bannon E, Ashrafioun L, Storfer-Isser A, Darby LA, Gumble A (2014) A randomized trial comparing two approaches to weight loss: Differences in weight loss maintenance. *J. Health Psych.* 19(2):296–311.

Charness G, Gneezy U (2009) Incentives to exercise. *Econometrica* 77(3):909–931.

Gertler P, Heckman J, Pinto R, Zanolini A, Vermeersch C, Walker S, Chang SM, Grantham-McGregor S (2014) Labor market returns to an early childhood stimulation intervention in Jamaica. *Science* 344(6187):998–1001.

Hussam R, Rabbani A, Reggiani G, Rigol N (2017) Habit formation and rational addiction: A field experiment in handwashing. Working paper, Harvard Business School, Boston.

Johnson EJ, Goldstein D (2003) Do defaults save lives? *Science* 302(5649): 1338–1339.

Jones D, Molitor D, Reif J (2019) What do workplace wellness programs do? Evidence from the Illinois Workplace Wellness Study. *Quart. J. Econom.* 134(4):1747–1791.

Judah G, Gardner B, Aunger R (2013) Forming a flossing habit: An exploratory study of the psychological determinants of habit formation. *British J. Health Psych.* 18(2):338–353.

Kuh GD, Kinzie JL, Buckley JA, Bridges BK, Hayek JC (2006) *What Matters to Student Success: A Review of the Literature*, vol. 8 (National Postsecondary Education Cooperative, Washington, DC).

Lally P, Chipperfield A, Wardle J (2008) Healthy habits: Efficacy of simple advice on weight control based on a habit-formation model. *Internat. J. Obesity* 32(4):700–707.

Larcom S, Rauch F, Willems T (2017) The benefits of forced experimentation: Striking evidence from the London Underground Network. *Quart. J. Econom.* 132(4):2019–2055.

Larrick RP, Soll JB (2008) The MPG illusion. *Science* 320(5883): 1593–1594.

Loewenstein G, Price J, Volpp K (2016) Habit formation in children: Evidence from incentives for healthy eating. *J. Health Econom.* 45:47–54.

Madrian BC, Shea DF (2001) The power of suggestion: Inertia in 401(k) participation and savings behavior. *Quart. J. Econom.* 116(4):1149–1187.

Mattke S, Schnyer C, Van Busum KR (2012) A review of the U.S. workplace wellness market. Occasional paper, RAND Corporation, Santa Monica, CA.

Milkman KL, Minson JA, Volpp KGM (2014) Holding the *Hunger Games* hostage at the gym: An evaluation of temptation bundling. *Management Sci.* 60(2):283–299.

Mokdad AH, Marks JS, Stroup DF, Gerberding JL (2004) Actual causes of death in the United States, 2000. *JAMA* 291(10): 1238–1245.

Neal DT, Wood W, Wu M, Kurlander D (2011) The pull of the past: When do habits persist despite conflict with motives? *Personality Soc. Psych. Bull.* 37(11):1428–1437.

Patel MS, Asch DA, Rosin R, Small DS, Bellamy SL, Heuer J, Sproat S, Hyson C, Haff N, Lee SM, Wesby L, Hoffer K, Shuttleworth D, Taylor DH, Hilbert V, Zhu J, Yang L, Wang X, Volpp KG (2016) Framing financial incentives to increase physical activity among overweight and obese adults: A randomized, controlled trial. *Ann. Internal Medicine* 164(6):385–394.

Royer H, Stehr M, Sydnor J (2015) Incentives, commitments, and habit formation in exercise: Evidence from a field experiment with workers at a Fortune-500 company. *Amer. Econom. J. Appl. Econom.* 7(3):51–84.

Schroeder SA (2007) We can do better—improving the health of the American people. *New England J. Medicine* 357(12):1221–1228.

Sen AP, Sewell TB, Riley EB, Stearman B, Bellamy SL, Hu MF, Tao Y, Zhu J, Park JD, Loewenstein G, Asch DA, Volpp KG (2014) Financial incentives for home-based health monitoring: A randomized controlled trial. *J. General Internal Medicine* 29(5): 770–777.

Staats BR, Dai H, Hofmann D, Milkman KL (2017) Motivating process compliance through individual electronic monitoring: An empirical examination of hand hygiene in healthcare. *Management Sci.* 63(5):1563–1585.

Tappe K, Tarves E, Oltarzewski J, Frum D (2013) Habit formation among regular exercisers at fitness centers: An exploratory study. *J. Physical Activity Health* 10(4):607–613.

Thaler RH, Benartzi S (2004) Save More tomorrow: Using behavioral economics to increase employee saving. *J. Political Econom.* 112(S1): S164–S187.

Thaler RH, Sunstein C (2008) *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Yale University Press, New Haven, CT).

Wood W, Neal DT (2016) Healthy through habit: Interventions for initiating & maintaining health behavior change. *Behav. Sci. Policy* 2(1):71–83.

Wood W, Rünger D (2016) Psychology of habit. *Annual Rev. Psych.* 67:289–314.