Creating Exercise Habits Using Incentives: The Tradeoff between Flexibility and Routinization

John Beshears
Harvard Business School

Hae Nim Lee
The Wharton School

Katherine L. Milkman
The Wharton School

Rob Mislavsky
The Wharton School

Last Modified: August 11, 2017

ABSTRACT: How can the formation of beneficial, lasting habits be promoted? Previous research suggests that persistent habits often involve regular, cue-triggered routines. We conducted a field experiment with 2,508 employees of a Fortune 500 company to test whether incentives for exercise routines—paying participants each time they visit a company gym within a daily two-hour window—lead to more persistent exercise behavior than flexible exercise incentives—paying participants each time they visit a company gym, regardless of the time of day. We find that an incremental gym visit in the daily two-hour window, compared to an incremental gym visit outside the window, was actually less likely to generate gym visits during the weeks after incentives were removed. Thus, while routines may be a common and important component of many lasting habits, encouraging overly rigid routines can undermine habit formation.

ACKNOWLEDGEMENTS: We thank Andrew Joung, David Mao, Byron Perpetua, and Predrag Pandiloski for outstanding research assistance. We also thank our partner organization for integral support on the design and execution of this experiment. Finally, we are grateful for the outstanding feedback we received on this work from seminar audiences at Carnegie Mellon University, Columbia University, Harvard University, Northwestern University, the University of Pennsylvania, and Yale University as well as 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, and the Behavior Change for Good Conference.

INTRODUCTION

Small, repeated, everyday decisions can have profound effects on many critical life outcomes. Choices that may seem trivial in the moment, such as what to eat, how much to exercise, how hard to study, and how to spend money, tend to accumulate over time to have large consequences (e.g., Kuh et al., 2006; Mokdad et al., 2004; Schroeder, 2007). For instance, adherence to a healthy lifestyle matters as much as genetic risk in predicting coronary artery disease—the leading cause of death worldwide (Khera et al., 2016). Policy interventions capable of shifting the habits that govern many everyday behaviors could improve welfare tremendously if applied to decisions about health, education, and personal finance (e.g., Beshears et al., 2013; Gertler et al., 2014; Loewenstein, Price, & Volpp, 2016).[1]

Psychology research has shown that stable habits tend to be characterized by engagement in behaviors under consistent circumstances or "routine" conditions—at the same time, in the same place, and following the same cue to act (Brooks et al., 2014; Tappe et al., 2013; Wood & Neal, 2016). This suggests an opportunity for policy makers: it may be possible to capitalize on the finding that routines are central to habitual behavior when attempting to encourage the formation of beneficial habits. Building on the past success of efforts to deploy psychological insights to improve the efficacy of policies seeking to shift behavior (Johnson & Goldstein, 2003; Larrick & Soll, 2008; Madrian & Shea, 2001),[2] we conducted a field experiment to determine whether people can be nudged to form longer-lasting exercise habits if they are encouraged to maintain a strict, routine schedule rather than to perform an activity frequently without any particular schedule. Consider two people hoping to establish a new, lasting exercise habit who begin a month-long program to help them reach this goal. One is encouraged to always visit the gym at a consistent time that she deemed convenient at the outset of the month (e.g., at 7:00 am), while the other is simply encouraged to visit the gym at any time without regard to building a stable routine. Our research answers the question: which person's exercise program leads to incremental gym visits that produce stronger habit formation? We evaluate whether an incremental gym visit induced at a routine time leads to more subsequent exercise in the long run than an incremental gym visit at any other time outside of that routine window.

The field experiment included 2,508 employees at a Fortune 500 company who were interested in exercising more regularly at their company's on-site gym. Participants were all exposed to a four-week intervention and were randomly assigned to one of five experimental conditions. At the beginning of the intervention, 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. Control group participants received no monetary incentives for exercise. Participants in two "flexible" experimental conditions were paid $3 and $7, respectively, for any weekday when they exercised for at least 30 minutes at their company's gym. Participants in two "routine" experimental conditions were also paid $3 and $7, respectively, for these workouts but *only* if they entered the gym within their chosen two-hour window. Our outcome of interest was how often participants visited the gym *after* the four-week intervention period, when incentive payments were no longer in place.

[1] For further references, see Milkman, Minson, & Volpp, 2014; Patel et al., 2016; Sen et al., 2014; Staats et al., forthcoming; and Thaler & Benartzi, 2004.

[2] For further reference see Thaler & Benartzi, 2004; Thaler & Sunstein, 2008; and Benartzi et al., forthcoming.

Recent research has shown that rewarding repeated engagement in desirable behaviors like exercise for as little as a month can build habits that stay in place long after incentives are removed (Acland & Levy, 2015; Charness & Gneezy, 2009; Royer, Stehr, & Sydnor, 2015). These findings are consistent with prior work theorizing and demonstrating that habits are formed by repeatedly engaging in a behavior (Becker & Murphy, 1988; Gruber & Koszegi, 2001, 2004; Gruber 2001; Wood & Neal, 2016), associating stable cues with that behavior (Brooks et al., 2014; Tappe et al., 2013; Wood & Neal, 2016), and linking a reward with the behavior (Wood & Neal, 2016). Thus, repeatedly rewarding desirable behaviors for a limited time period may be a useful policy for promoting habit formation (if the policy is cost-effective when applied to a given setting). In this paper, we address whether an insight from psychology about the importance of routines to habit formation makes it possible to improve upon the "rewards for repeated behavior" approach.

Specifically, a growing body of evidence suggests that creating stable, cued routines may be critical to successfully promoting habit formation. For example, forming if-then plans and committing to take a given action after a certain cue (e.g., a reminder, a specific time of day) increases engagement in desirable behaviors (Gollwitzer, 1999; Milkman et al., 2011; Rogers et al., 2015). Correlational studies have also shown that people with cue-triggered routines engage in desirable behaviors more regularly than others. Approximately 90% of regular exercisers associate a location or time cue with exercise, and exercising is more automatic for people with cues that trigger exercise (Tappe et al., 2013). In the domain of medication regimens, adherence is higher among those with regular, pill-taking routines (Brooks et al., 2014). In addition, "piggybacking" habit formation interventions that encourage people to add a desirable behavior to a routine they already engage in (e.g., adding flossing to an existing tooth-brushing routine) are more effective than interventions that simply encourage a new, desirable behavior (Phillips, Leventhal, & Leventhal, 2013; Judah, Gardner, & Aunger, 2013). Together, these findings suggest that incentivizing the creation of stable, cue-based routines may be a potent strategy for generating lasting habits. A survey that we conducted of 69 psychology professors at top 40 universities as ranked by *U.S. News and World Report* (2016) suggests widespread agreement with this proposition. Specifically, when we posed our research question, asking whether an incremental gym visit induced at a routine time would lead to more subsequent exercise in the long run than an incremental gym visit at any other time, 77% of our expert sample predicted that an incremental gym visit induced at a routine time would produce more subsequent exercise (see Appendix G for full study details).

However, past research and theory also point to the value of flexibility. Given a choice between receiving a reward for enacting a behavior (a) at any time or (b) during a narrow, self-determined window, a classical economic agent would prefer option (a) because it weakly increases the probability of receiving the reward. Goals for increasing engagement in desirable behaviors are easier to achieve with flexible rewards, and when attempting to promote long-term behavior change, small failures to achieve goals can be harmful, causing people to reduce their overall performance (Cochran & Tesser, 1996; Polivy, 1976; Sharif & Shu, in press; Soman & Cheema, 2004).

Our field experiment is designed to test whether incremental incentive-induced routine gym visits—visits that begin during the chosen two-hour window—lead to more exercise after incentives are removed than incremental incentive-induced non-routine gym visits—visits that begin outside the chosen two-hour window. The five experimental conditions generate the variation necessary to address this question. The flexible conditions, which offered $3 or $7 for

each day a participant visited the gym regardless of the time of day, led to increases in the number of in-window and out-of-window gym visits during the incentive period. The $3 flexible condition increased in-window visits by 0.32 per week and out-of-window visits by 0.26 per week relative to the control condition, and the $7 flexible condition increased in-window visits by 0.43 per week and out-of-window visits by 0.46 per week relative to the control condition. The routine conditions, which offered $3 or $7 only if the gym visit began during the chosen two-hour window, led to increases in the number of in-window gym visits and decreases in the number of out-of-window gym visits during the incentive period. The $3 routine condition increased in-window visits by 0.58 per week and decreased out-of-window visits by 0.18 per week relative to the control condition, and the $7 routine condition increased in-window visits by 0.97 per week and decreased out-of-window visits by 0.27 per week relative to the control condition.

The key outcome of interest, of course, is exercise behavior after the four-week incentive period ended. However, we do not focus on comparing the absolute number of gym visits after the incentive period across the five experimental conditions. Because the flexible conditions exhibited higher levels of exercise overall during the incentive period compared to the routine conditions, our finding that the flexible conditions also exhibited higher levels of exercise after the incentive period compared to the routine conditions does not necessarily indicate that in-window versus out-of-window gym visits during the incentive period have differential consequences for the formation of exercise habits. It may simply be the case that more exercise in the past begets more exercise in the future, regardless of the timing of the past gym visits. Instead of comparing the absolute number of gym visits after the incentive period across the experimental conditions, we would like to compare one incremental in-window visit during the incentive period to one incremental out-of-window visit during the incentive period, measuring the impacts of those incremental visits on subsequent exercise.

To address this question, we use an instrumental variables regression framework. The outcome variable in these regressions is the total number of gym visits per week during the four weeks immediately following the incentive 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 incentive period. We also conduct versions of the analysis that focus only on in-window visits or only on out-of-window visits during the four weeks following the incentive 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 during the incentive period, and we instrument for these two variables using four treatment group indicators, omitting an indicator for the control group. Using this framework, we find that each incremental weekly in-window gym visit during the incentive period leads to 0.25 extra total gym visits per week and an 11 percentage point increase in the likelihood of visiting the gym at least once in a given week during the following four weeks. Each incremental weekly out-of-window gym visit during the incentive period leads to 0.33 extra total gym visits per week and a 17 percentage point increase in the likelihood of visiting the gym at least once in a given week during the following four weeks. The difference between the coefficients is statistically significant when the outcome variable is the indicator for visiting the gym at least once in a given week. Thus, the results indicate that incentive-induced routine gym visits are less effective than incentive-induced non-routine gym visits when it comes to generating sustained exercise habits in the future.

This paper fills a gap in our understanding of how habits are formed and how to promote habit formation by examining whether encouraging routinized, cue-based routines or flexible

repetition of a desirable behavior offers a superior path to habit formation. Further, it does so in a critically important context. Recent research points to poor physical fitness as the second most important risk factor for premature mortality, falling only behind smoking in importance (Ladenvall et al., 2016), and poor diet and physical inactivity account for approximately 400,000 premature deaths in the U.S. each year (Mokdad et al., 2004). Although exercise promotes weight loss (Andersen, 1999) – a particularly critical outcome given that 68% of Americans were overweight as of 2008 (Flegal et al., 2010) – it also has unrelated health benefits (Lee, Blair, & Jackson, 1999) such as improving self-esteem (Brownell, 1995) and reducing depression, stress, and anxiety (Brownell, 1995; Kayman, Bruvold, & Stern, 1990). However, despite the many benefits of physical activity (Andersen, 1999; Ding et al., 2016), only 50% of Americans exercise sufficiently (CDC, 2007).

In this paper, we study the efficacy of a new approach to promoting long-term habit formation in the context of exercise. Our research therefore contributes to basic theories of habit formation as well as offering immediately policy-relevant insights about how to promote increased physical activity and increased engagement in other important, repeated behaviors. Recent research has highlighted growing interest not only from academics but also from employers in finding effective ways to encouraging healthy habits (Gallani 2017). Through our 2,500-person field experiment, we test whether incentivizing routinized or flexible exercise habits is best for habit formation and show that incentive-induced routine gym visits create less persistent habits than incentive-induced non-routine gym visits. We also present evidence that small ($3-7) repeated rewards for exercise can produce detectable changes in behavior that last for up to forty weeks after a 4-week intervention period.

The remainder of this paper is organized as follows. Section 2 presents the design of our experiment and methods for analyzing resulting data. Section 3 presents our experimental results. Section 4 discusses our findings and their interpretation, and Section 5 concludes.

2. EXPERIMENTAL DESIGN AND IMPLEMENTATION

Setting

We collaborated with a large technology company to conduct a randomized controlled trial with a subset of the company's employees. To be eligible to participate, an employee was required to be a full-time, part-time, or fixed-term employee or intern at one of the company's seven included U.S. office locations, 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 are equipped with exercise machines and weights. Gym membership is free to all employees. Upon entering the gym, employees encounter a computer monitor where they are asked to record their gym visit by swiping their employee identification badge. We rely on these login data to track individual gym attendance. Employees are also asked to swipe their badge as they exit the gym.

Participant Recruitment and Randomization

Figure 1 shows the flow and randomization of study participants, and Figure 2 illustrates the timeline of the experiment.

Recruitment. Participant recruitment began on February 3, 2015, through a series of poster and email advertisements (see Appendix A, Figures A1 and A2). 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.

Registration Survey. Employees who responded to our recruitment campaign were given a web link to complete our registration survey (Appendix A, Figures A10). Upon starting the survey, employees were told that the Fresh Start Fitness Challenge was part of a research study being conducted by their employer 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 over-enrollment.

The survey began with a consent form and some background questions (name, email address, office location, typical number of days per week they exercised for at least 30 minutes, gender, and ethnicity). Next, employees were asked to register their employee identification badge with the company gym, allowing us to track their gym entrances and exits (see 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 experiment) 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 Appendix B for more detailed information about the partner pairing system).

After choosing a workout partner, employees were asked to select a two-hour block of time when they preferred to start their weekday workouts (of at least 30 minutes) at the company gym.[3] Although employees could coordinate workout windows with their partners (31.7% of the final sample selected a workout window that overlapped perfectly[4] 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 window. They could also opt-in to 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

[3] 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.

[4] In our initial survey 117 individuals had missing observations for their window selections. We define perfect overlaps with partner's window as a binary variable coded as 1 if workout partners have a perfectly matching (identical) selection of workout windows or if they both have a missing selection in the initial survey. This definition was used for the randomization process and the construction of our workout window control variable. However, we manually imputed selection of workout windows for 116 of the 117 individuals after the randomization. We summarize the complete variable in Table 2 and use it to construct the in-window and out-of-window outcome variables.

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 Appendix A, Figure A5). 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).[5] Out 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.

Experimental Conditions. Each participating pair of employees was randomly assigned to one of five conditions (four treatment conditions and one control condition), with randomization occurring by pair. 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, where participants earned a payment for each weekday (Monday-Friday) during which they worked out at the company gym for at least 30 minutes. The other two treatment conditions were *routine* conditions, where 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.

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 (Acland & Levy, 2015; Charness & Gneezy, 2009; Royer, Stehr, & Sydnor, 2015). Incentives of approximately the same magnitude as ours increased the number of gym visits per week by 33%-213%, with a median 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

[5] 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, as they were not variables of interest to our team.

80% power to detect a 15%-20% difference in gym visits between the *flexible* and *routine* conditions, which required at least 750 participants per condition.[6]

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 and 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, however, why our recruitment materials and intake survey encouraged employees to sign up for the study with a workout partner.

Randomization. Our registration survey closed on February 25, 2015 (two days later than initially planned, as 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, (b) registered online with the company gym, and (c) had a partner who had registered online with the company gym.[7] 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 or not the partners had (spontaneously) coordinated their workout windows. The randomization scheme therefore had 12 strata total, four for each of the three randomization waves. All regression results that we report control for strata fixed effects.

---

[6] 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), which gave us 80% power to detect a 34% 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 post-experiment using the observed intracluster correlation in our sample of 0.26, we determined that the detectable effect sizes in our study were 38% and 16%, respectively.

[7] 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 upon entry, 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 Appendix A, Figure A6).

The Intervention

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 (see Appendix A, Figure A7) containing links to a website describing the incentive structure for his or her condition and to a comprehension check survey (see Appendix A, Figure A14). 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 people other than their workout buddy about the Fresh Start Fitness Challenge. However, we could not monitor or enforce compliance with this request.

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 (Appendix F contains the exact contents of the reminder messages).[8]

Post-Intervention Period. On April 17, 2015 (two weeks after the intervention period ended), participants received an email (see Appendix A, Figure A8) asking them to complete an exit survey (see Appendix E). After the exit survey was completed, participants received study-related payments through an online payment system. During the post-intervention period, we continued to collect gym attendance data. In addition, participants continued to receive daily workout reminders for 10 months post-intervention (until February 1, 2016) unless they opted out.

To encourage participants to continue to reliably swipe their employee identification badges when entering and exiting the gym, we announced on April 1 (the first day after the conclusion of the intervention period) that we would be conducting a monthly lottery through the end of 2015. A randomly selected participant would receive a $250 Visa gift card if that participant swiped in *and* out of the gym on a randomly selected day during the month in question (see Appendix A, Figure A9).

Statistical Analysis

Our primary outcome variable 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 (Acland & Levy, 2015; Charness & Gneezy, 2009; Milkman et al., 2014), we initially planned to obtain and analyze data from two post-intervention follow-up periods: (1) the four-week period following the conclusion of the intervention (a length of time mirroring the length of our intervention) and (2) the ten-week period following the conclusion of the intervention (mapping roughly onto the follow-up periods from Charness and Gneezy, 2009 Study 1; Acland and Levy, 2015; and Milkman et al., 2014). However, we were unexpectedly also able to obtain data through the end

[8] 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.

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 post-intervention period, but analogous analyses for the ten-week and 40-week post-intervention periods can be found in the Appendix Tables and Figures.

The primary objective of the experiment is to evaluate the impact of inducing additional gym visits during participants' routine "workout windows" on habit formation compared with the impact of inducing additional gym visits at other, arbitrary times. To facilitate this comparison, we rely on an instrumental variables framework to analyze our experimental data.

A gym visit is considered an "in-window gym visit" if it begins during the participant's pre-selected two-hour workout window (e.g., between 1:00 pm and 3:00 pm if the participant chose 1:00pm-3:00pm as his or her preferred workout window during the registration survey). A gym visit is considered an "out-of-window gym visit" if it begins at any other time. In the first stage of the instrumental variables analysis, we use the experimental conditions to which participants are assigned as instruments to predict in-window and out-of-window gym attendance during the intervention period. In the second stage, we use predicted in-window and out-of-window gym attendance during the intervention period to predict in-window, out-of-window, and total gym attendance post-intervention.

The instrumental variables analysis facilitates an interpretation of the experimental results that is more convenient than the interpretation of a comparison of average post-intervention gym attendance across experimental conditions. For example, participants in the *flexible $7* payment group might exhibit higher average levels of post-intervention gym attendance than participants in the *routine $7* payment group, but this pattern could be driven by (a) greater persistence of out-of-window exercise behaviors (which are more common in the *flexible $7* payment group) compared to in-window exercise behaviors (which are more common in the *routine $7* payment group) or (b) more frequent gym attendance overall (whether in-window or out-of-window) during the intervention period among participants in the *flexible $7* payment group than among participants in the *routine $7* payment group, which leads to higher levels of post-intervention gym attendance in the former group than in the latter group because of the difference in the stock of accumulated past exercise behavior (Becker and Murphy, 1988). The instrumental variables strategy is designed to set aside the second factor and focus on the first factor—the persistence of in-window versus out-of-window exercise behaviors—by scaling the treatment effect estimates so that they indicate the impact of one incremental in-window gym visit per week during the intervention and the impact of one incremental out-of-window gym visit per week during the intervention on post-intervention exercise behaviors. Furthermore, the instrumental variables strategy sidesteps the influence of omitted factors such as individual-level heterogeneity in the propensity to exercise by using only the component of variation in the frequency of in-window and out-of-window gym attendance during the intervention that is driven by randomly assigned experimental conditions.

Dependent Variables. 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 in each week with a gym visit made by each of our study participants (e.g., Acland & Levy, 2015; Charness & Gneezy, 2009; Milkman et al., 2014; Royer et al., 2015). Second, we measure whether or not 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 1 if a participant

visited the gym at least once during the week and 0 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.[9] We use analogous definitions to calculate the total number of in-window gym visits and the total number of out-of-window gym visits made during each week by each of our study participants, as well as the binary variables capturing whether or not a participant made an in-window gym visit and whether or not a participant made an out-of-window gym visit at least once in a given week. Particularly, we calculate the total number of in-window gym visits as the number of days in a given week with an in-window visit, and we calculate the total number of out-of-window gym visits as the number of days in a given week in which study participants make an out-of-window gym visit, *but not* an in-window gym visit. In other words, if study participants make both an in-window and out-of-window gym visit on the same day, we count that day only toward their total number of in-window gym visits.

Regression Specifications. As described above, to measure the effect of inducing one additional in-window gym visit per week and the effect of inducing one additional out-of-window gym visit per week during the intervention period on post-intervention exercise behavior, we use an instrumental variables framework.

In the first stage, we predict a participant's number of in-window gym visits per week and number of out-of-window gym visits per week during the intervention period. The regression specification is:

$$W_i^j = \alpha_{j0} + \alpha_{j1} C_{Flex\ \$3,i} + \alpha_{j2} C_{Flex\ \$7,i} + \alpha_{j3} C_{Rout\ \$3,i} + \alpha_{j4} C_{Rout\ \$7,i} + \boldsymbol{\beta_j' X_i} + \varepsilon_{ji} \qquad (1)$$

where *i* indexes participants, *j* indicates whether the equation concerns in-window or out-of-window gym visits, the left-hand-side variable is the number of in-window gym visits per week during the intervention period (W_i^{IN}) or the number of out-of-window gym visits per week during the intervention period (W_i^{OUT}), the instruments 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 (1) strata in our randomization scheme, which 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 or not the partners had (spontaneously) coordinated their workout windows, as well as (2) twelve interaction effects between these three variables on which we stratified random assignment to condition.

In the second stage, we predict outcome variables measuring gym attendance during the post-incentive period using the predicted number of in-window gym visits per week during the intervention period ($\widehat{W}_i^{IN}$) and the predicted number of out-of-window gym visits per week during the intervention period ($\widehat{W}_i^{OUT}$) from the first stage. The regression specification is:

$$y_{it} = \gamma_0 + \gamma_1 \widehat{W}_i^{IN} + \gamma_2 \widehat{W}_i^{OUT} + \boldsymbol{\delta' X_i} + \zeta_{it} \qquad (2)$$

[9] Note that to earn incentives 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.

where *i* again indexes participants, *t* indexes weeks, and $\boldsymbol{X_i}$ is the same vector of control variables from the first stage. The left-hand-side variable y_{it} is one of six outcomes:

1. Total number of gym visits for participant *i* during week *t*
2. Number of in-window gym visits for participant *i* during week *t*
3. Number of out-of-window gym visits 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/her workout window during week *t*
6. Whether participant *i* visited the gym outside of his/her workout window during week *t*

For ease of interpretation, we use linear models in the main tables of the paper, even when the outcome variable is binary. However, we also provide the results of instrumental variables probit models in Appendix Table 5, and these regressions yield the same results. Throughout the analysis, we cluster standard errors at the participant pair level.

Discussion of Estimation Framework. The assumptions necessary for the instrumental variables strategy to be valid are likely to be satisfied. The exogeneity of the instruments is a feature of the experimental design—the treatment group indicators are randomly assigned and therefore orthogonal to individual characteristics and environmental factors that may cause some employees to visit the gym more or less frequently. The exclusion restriction, while ultimately untestable, seems reasonable. It is intuitively plausible that the channel by which treatment group assignment influences exercise behavior after the incentive period is through repeated engagement in exercise behavior during the incentive period. Alternative channels, such as income effects created by incentive payments or informational effects having to do with employee inferences regarding the strength of the employer's exercise recommendations, seem less likely as explanations for the results, although we cannot rule them out. Perhaps the most questionable assumption behind the instrumental variables strategy is the linear structure we have imposed by specifying that post-intervention exercise behavior is a linear function of in-window gym visits during the incentive period and out-of-window gym visits during the incentive period. However, we assess this assumption in our robustness checks by testing more flexible functional forms, and we find that the linear structure is a good approximation.

3. RESULTS

Data

Table 1 presents summary statistics for the self-reported variables collected in our pre-intervention registration survey: age, company tenure, weekly pre-intervention workout frequency, BMI (calculated from self-reported height and weight), gender, job function, job level, and office location. This table shows the means, standard deviations, and the proportion of participants who responded to each question for all participants in our study (Column 1) as well as participants in our *control* group (Column 2), *flexible* groups (Columns 3-5), and *routine* groups (Columns 6-8). Performing pairwise t-tests to compare each demographic variable across experimental conditions, we find that only 8 out of the 100 possible comparisons differ significantly at an alpha level of 0.05, or roughly the number (5) that would be expected by

chance. Thus it appears that random assignment successfully achieved balance across conditions.

Treatment Effects During the Intervention Period

Main Effects of Interventions on Exercise. Patterns of gym attendance over the four-week incentivized intervention period depicted in Figure 3 suggest that our incentives worked during this period as expected such that larger payments yielded more exercise, and routine incentives yielded more in-window workouts but fewer overall workouts. Figure 3 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, and Table 3 presents our first stage IV results comparing overall, in-window and out-of-window gym attendance across treatment conditions.

Incentive Size. Overall, participants in our incentivized conditions completed an average of 1.75 weekly workouts (SD=1.66) and 66.8% completed one workout or more per week during our intervention period, on average. These incentivized participants exercised significantly more during the intervention period than participants in our control condition who only completed an average of 1.11 weekly workouts (SD=1.47) and only 49.8% of whom completed one or more workouts per week, on average, during our intervention period (both p's < 0.001). Higher incentive payments also lead to more exercise during our intervention period. As Table 3A, Column 4 depicts, participants paid \$7 visited the gym a regression-estimated 0.30 more times per week than those paid \$3 during the intervention period ($p<0.001$); and as Table 3B, Column 4 depicts, participants paid \$7 visited the gym one or more times at a regression-estimated six percentage point higher rate during the incentive period than participants paid \$3 ($p<0.001$).

Flexible vs. Routine Incentives. As Table 3A Column 7 depicts, 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 3B Column 7 depicts, participants in the *flexible* conditions visited the gym one or more times at a regression-estimated four percentage point higher rate during the incentive period than participants in the *routine* conditions ($p<0.05$). These results are not surprising because our *flexible* condition made it easier to earn incentives for exercise. The reason we varied incentive payments within our flexible and routine conditions was in fact to induce additional variation in weekly workouts above and beyond that produced by changing the flexibility of incentive payments so we could conduct our instrumental variables analysis to measure the impact of inducing extra in-window vs. out-of-window workouts on habit formation.

As expected, and confirming the success of our manipulation, during our intervention period participants in the *routine* conditions exercised significantly more during their workout windows (completing 1.36 weekly in-window workouts on average (SD=0.09) and at least one in-window workout per week 56.5% of the time, on average) than did participants in the *flexible* conditions (who completed 0.97 weekly in-window workouts on average (SD=0.09) and at least one in-window workout per week 46.7% of the time, on average; both Wald tests reported in Tables 3A-B, Column 8: p's < 0.001). Conversely, those in the *flexible* conditions exercised significantly more outside of their workout windows (completing 0.88 weekly out-of-window workouts on average (SD=0.09) and at least one out-of-window workout per week 56.4% of the

time, on average than did participants in the *routine* conditions (who completed 0.30 weekly out-of-window workouts on average (SD=0.08) and at least one out-of-window workout per week 20.4% of the time, on average; both Wald tests reported in Tables 3A-B, Column 9: p's < 0.001).

Another way of validating the success of our manipulation is to examine the fraction of participants' workouts that took place during their workout windows by experimental condition. We find that 76.9% of participants' workouts in the *routine* conditions took place in-window, a significantly higher fraction than (a) the 50.4% of workouts in the *flexible* conditions that were completed in-window ($p<0.001$) and (b) the 52.3% of workouts in the *control* condition that were completed in-window ($p<0.001$).

Post-Intervention Results

Patterns of post-intervention gym attendance over our four-week follow-up period are depicted in Figure 4. Specifically, Figure 4 presents means of weekly overall, in-window and out-of-window gym attendance by experimental condition post-intervention during our four-week follow-up period. Table 4 presents regressions comparing overall, in-window and out-of-window gym attendance across treatment conditions, and Table 5 presents our key second stage instrumental variables regression results comparing the impact of inducing additional in-window and out-of-window workouts on additional overall, in-window and out-of-window workouts post-intervention.

Results by Condition. Average regression-estimated post-intervention gym attendance during our 4-week follow-up period by condition can be found in Tables 4A (which focuses on predicting participants' number of weekly workouts) and 4B (which focuses on predicting participants' likelihood of working out during a given week). First, as Tables 4A-B show, we replicate the well-established findings from Charness and Gneezy (2009), Acland and Levy (2015) and Royer et al. (2015) that when participants are paid to exercise any time of day for a month, even after payments conclude, paid participants continue to exercise significantly more during our 4-week follow-up period (approximately 33% more; see Table 4A Column 7) than those in an unpaid control group. On average, as Table 4A, Column 7 reports, participants in the *flexible* conditions exercised more, overall, in the 4-week post-intervention period (completing an average of 1.01 workouts per week, SD=0.11) than those in either the *routine* conditions (who completed an average of 0.91 workouts per week, SD=0.11; $p<0.05$) or the *control* condition (who completed an average of 0.76 workouts per week, SD=1.32; $p<0.05$). As depicted in Table 4bB Column 7, a similar pattern emerges when we consider a participant's likelihood of working out in a given week. Participants in the *flexible* conditions were more likely to visit the gym at least once during our follow-up period (45.6% made one or more weekly visits, on average) than those in either the *routine* conditions (39.9% of whom made one or more gym visits, on average; $p < 0.001$) or the *control* condition (33.9% of whom visited the gym at least once, on average; $p < 0.001$).

Interestingly, although participants in the *routine* conditions worked out less frequently overall post-intervention than those in the *flexible* conditions, they worked out very slightly (though insignificantly) more often during their workout windows (0.53 times per week, on average; SD=0.07) than those in the *flexible* conditions (0.50 times per week, on average; SD=0.07; see Table 4A Column 8). As depicted in Figure 4B, the group with the most in-window workouts post-intervention was the *routine $7* payment group. Similarly, although

participants in the *routine* conditions were less likely to make any weekly visits to the gym post-intervention than participants in the *flexible* conditions, as Table 4B Column 8 shows (and Figure 4E depicts), participants in the *flexible* conditions worked out at least once during their workout windows (27.6% made 1+ weekly visits, on average) at almost exactly the same rate as participants in the routine conditions (27.5% made 1+ weekly visits, on average).[10] These findings offer some support for the possibility that our *routine* condition successfully encouraged routine formation such that participants in these conditions developed the habit of visiting the gym during their workout windows. We will rely on instrumental variables analyses to more carefully explore this hypothesis in a moment.

Before turning to our instrumental variables analysis, it is worth noting what will now not come as a surprise since overall workouts are the sum of in-window and out-of-window workouts: the reason participants in the *flexible* conditions workout more overall than participants in the *routine* conditions is that they are much more likely to workout outside of their workout windows post-intervention (see Tables 4A-B, Column 9 and see Figures 4C and 4F). Specifically, participants in the *flexible* conditions complete an average of 0.51 out-of-window workouts per week post-intervention (SD=0.06), significantly more than participants in the *routine* conditions who completed 0.38 such workouts per week post-intervention, on average (SD=0.06; $p < 0.001$). Similarly, participants in the *flexible* conditions were also more likely to make any visits to the gym (29.6% of the time, on average) than participants in the *routine* conditions (who visited one or more times weekly 22.9% of the time, on average; $p < 0.001$). Participants in the *routine* conditions, in fact, exercised outside of their workout windows at similar rates to participants in the *control* condition (who completed 0.34 out-of-window workouts per week, on average post-intervention and 21.2% of whom visited the gym one or more times a week out-of-window, on average), suggesting that the only lasting, post-intervention exercise habits induced for the *routine* participants corresponded to exercise during their in-window workout times.

IV Analysis: Effects of Inducing Additional In-window vs. Out-of-window Workouts. Our central research question can be addressed with instrumental variables analysis. We are fundamentally interested in determining how (experimentally) inducing additional in-window versus out-of-window workouts during our intervention period affects post-intervention exercise overall, in-window and out-of-window. Table 5 presents the results of our second stage instrumental variables regressions for our 4-week post-intervention period and addresses these questions. Specifically, this table uses the predicted number of in-window and out-of-window workouts induced by our treatment conditions during our intervention period (from our stage 1 regressions, see Table 3) to predict all post-intervention workouts (columns 1 and 2), in-window post-intervention workouts (columns 3 and 4), and out-of-window post-intervention workouts (columns 5 and 6). Specifically, we predict the number of weekly workouts (columns 1, 3, and 5) and the likelihood that a participant exercises in a given week post-intervention (columns 2, 4, and 6).

The coefficients reported in Table 5 provide an estimate of the impact of inducing one additional in-window or out-of-window workout during our intervention period on post-

[10] Unsurprisingly given that they exercised less overall, participants in our *control* condition also exercised slightly although insignificantly less often post-intervention, in-window, than participants in other conditions (completing 0.42 weekly workouts, on average, SD=0.95 and visiting the gym at least once a week 22.3% of the time, on average).

intervention exercise. Columns 1 and 2 of Table 5 show that inducing additional workouts *either* in-window or out-of-window during our intervention period significantly increase total post-intervention workouts (Column 1) and likelihood of making any weekly gym visits (Column 2). However, inducing additional out-of-window workouts produces a directionally (and sometimes significantly) larger effect than inducing additional in-window workouts (generating an extra 0.08 workouts per week, p=0.33; and a 6-percentage point larger increase in the likelihood of one or more weekly gym visits, on average, p<0.05). As Columns 5 and 6 of Table 5 show, this effect is driven by the large difference in the number of out-of-window workouts yielded during our 10-week post-intervention follow-up period by inducing additional out-of-window workouts during our four-week intervention. Specifically, as Table 5 Column 5 shows, inducing one additional out-of-window workout during our intervention period leads participants to visit the gym 0.20 more times per week out-of-window post-intervention than inducing one additional in-window workout (p <0.001), and as Table 5 Column 6 shows, inducing one additional out-of-window workout during our intervention period leads to a 10 percentage point larger increase in the number of participants making one or more out-of-window gym visits per week, on average, post-intervention than inducing one additional in-window workout during our intervention period ($p < 0.001$).

Consistent with the idea that routines beget routines, inducing one additional weekly in-window workout during our intervention period produces 0.12 more total in-window workouts post-intervention than inducing one additional weekly out-of-window workout during our intervention period ($p < 0.10$) as Table 5 Column 3 shows. And we see the same directional pattern of results (though the difference is insignificant) when we look instead at the outcome of inducing any in-window weekly workouts post-intervention (see Table 5 Column 4). However, these boosts in in-window workouts post-intervention induced by additional in-window workouts during the intervention period are swamped by the aforementioned benefits of inducing additional out-of-window workouts during our intervention period on post-intervention out-of-window workouts.

While raw experimental data can't be visually inspected to depict the results of an instrumental variables analysis, Figure 5 offers a schematic visual analog of our primary finding by focusing in on a subset of our experimental data. Specifically, Figure 5 zooms in on two experimental conditions that induced similar numbers of total workouts during our intervention period – *routine $7* payment and *flexible $3* payment. Notably, the workouts these experimental conditions induced during our intervention period were distributed differently: the *routine $7* payment condition produced many more in-window workouts and fewer out-of-window workouts than the *flexible $3* payment condition. As Figure 5 shows, while the *routine $7* payment condition generated slightly more workouts during our intervention period, the efficacy of these two conditions post-intervention flipped, with the *flexible $3* payment condition generating a directionally more lasting post-intervention pattern of workouts. Figure 5 reinforces our key finding: that inducing additional out-of-window workouts during our four-week intervention produces directionally (and sometimes significantly) more post-intervention exercise than inducing additional in-window workouts.

Heterogeneity Analyses. We conducted a number of heterogeneity analyses to determine whether our results varied as a function of individual characteristics such as pre-intervention exercise frequency, BMI, and so on. We found no significant heterogeneity in our treatment effects as a function of the time of day when a participant scheduled her workout window,

whether a participant chose to synch her workout window with her partner's, whether a participant was an above (vs. below) median exerciser pre-intervention (based on the self-reported typical number of pre-intervention workouts he/she completed weekly), participant BMI (based on self-reported weight and height), or participant job level. We also explored whether the flexibility of an employee's work schedule had any impact on our results by conducting heterogeneity analyses by participant job function (e.g., business and administration, engineering, etc.). We do this under the assumption that different job functions have differing levels of flexibility, which may impact the ability to complete in-window workouts. We found no significant heterogeneity by participant job function, however.

Longevity of Effects. We also examined the longevity of our findings. Specifically, we examined if the benefit of inducing one additional routine vs. flexible workout lasts. To do this, we simply repeat our instrumental variables analysis described above with data on participants' gym visits over the 10 and 40 week periods post-intervention (Appendix Tables 2 and 4, respectively).

As shown in Appendix Table 2, the benefits of inducing both one additional routine workout and one additional flexible workout during the intervention period remain positive and significant (or marginally significant in one case) on post-intervention exercise for a 10 week follow-up period, and induced out-of-window workouts continue to have comparably more enduring benefits than in-window workouts. When considering the entire 40 week follow-up period (see Appendix Table 4), again, the benefits of inducing both one additional routine workout and one additional flexible workout during the intervention period remain positive and significant (or marginally significant in one case) for post-intervention exercise. However, induced out-of-window workouts no longer have a significantly larger benefit than in-window workouts in any specification (though the difference remains directionally the same).

Robustness Checks. We present a number of analyses showing the robustness of our key results. First, Appendix Tables 2 and 4 show our results when we examine 10-week and 40-week post-intervention time periods, respectively. Unsurprisingly, our results weaken as our follow-up period lengthens, but the general conclusions and patterns of result are robust to these longer follow-up periods. We also present probit regressions to evaluate the dependent variable "likelihood of exercising in a given week" for all specifications in Appendix Tables 5A-C and show that our results from ordinary least squares regressions are robust to this alternative analysis strategy. When we re-analyze our data including more controls, we obtain largely the same pattern of results as well, as shown in Appendix Tables 6 to 11. These additional controls are indicators for (a) registration date; (b) age deciles; (c) deciles of tenure at the company; (d) the individual's typical number of workouts per week, as reported in the registration survey; (e) deciles of BMI; (f) gender; (g) job function; (h) job level; (i) office location; (j) ethnicity; (k) whether or not the individual signed up for text message workout window reminders in the registration survey;[11] (l) workout window time (morning, midday, afternoon, or evening); (m) whether or not the individual connected a Fitbit device to an AchieveMint account; and (n)

[11] The default setting was to receive email workout window reminders, and participants were only able to opt out of these email reminders after randomization took place.

having a missing value for a given item in this list. See Tables 1 and 2 for summary statistics describing these control variables.[12] As reported in Appendix Tables 12 to 18 we see the same basic pattern of results when we re-analyze data collected only in the largest office location of the corporation we partnered with in this experiment, though a number of our findings are no longer significant when we look only at a single office location because of our reduced statistical power. As reported in Appendix Tables 19-29, our results are also robust to examining the minutes a participants spent at the gym per week (assuming 30 minute visits for anyone who failed to badge out within 4 hours) as an alternative dependent variable or to only counting gym visits when participants badged out at least 30 minutes after badging in. Finally, in Appendix Table 41, we relax the assumption that post-intervention exercise is a linear function of in-window exercise during the intervention and out-of-window exercise during the intervention. When we add quadratic terms for in-window exercise during the intervention and out-of-window exercise during the intervention, we find little evidence of curvature, indicating that the linear function is a good approximation to the true relationship.

Validation of Data on Gym Usage

We took two primary steps to verify that participants were indeed exercising when they swiped into the gym and were equally likely to swipe into the gym when they visited across experimental conditions. First, a small subset of participants (N = 114) successfully registered their fitness trackers (primarily Fitbits) with AchieveMint, a website that rewards users for physical activity. We were able to retrieve minute-by-minute step data for these participants. As depicted in Appendix Figure 3, during our four-week intervention period, we observe a substantial increase in the number of steps taken in the 30 minutes after these participants swiped into the gym (M_{steps} = 2,190.35, SD_{steps} = 1775.13) relative to the average number of steps they took in 30 minute periods outside of the gym (M_{steps} = 210.44, SD_{steps} = 133.32; $p < 0.001$). We observe the same pattern post-intervention where participants take dramatically more steps during their 30 minutes after entering the gym than at other times (30 minutes post-swipe: M_{steps} = 1,900.63, SD_{steps} = 1896.29; other 30 minute intervals: M_{steps} = 192.31, SD_{steps} = 139.41 $p < 0.001$). Together these data (depicted in Appendix Figure 3) suggest that people are indeed exercising at the gym. Second, in our post-intervention survey (see Appendix E for complete questionnaire), we asked participants how likely they were to badge into the gym during a visit (rather than visiting but failing to badge in) over the previous week. We found that there were no differences between conditions in this measure, with responses ranging from 76.5% (SE = 2.61%) in the routine $7 condition to 81.4% (SE = 3.62%) in the control condition (all pairwise *p*'s > 0.26).[13]

[12] Note that Table 2 reports summary statistics as of the end of the intervention period, while the control variables used in these regression analysis are measured as of the date of randomization. The summary statistics as of the end of the intervention period are more interesting to display (for example, the percentage of participants receiving email reminders is 100% as of the date of randomization because of the study design), but the summary statistics as of the date of randomization are very similar.

[13] As another means of validating that employees reliably badge into the gym, a group of employees of our corporate partner who were not directly involved with our experiment performed audits on badging rates at two different company gyms on two different dates during our intervention period, observing all gym users, not just those in our

4. DISCUSSION

Interpretation of Results

Our experiment is not set up to isolate the mechanisms that cause in-window gym visits to generate less persistent exercise behavior than out-of-window gym visits, but we discuss several possible reasons for our results. First, it is possible that non-constant marginal returns from exercise are the reason why an incremental in-window gym visit leads to less subsequent exercise than an incremental out-of-window gym visit. However, if each incremental in-window gym visit were to have a lower or higher impact on subsequent exercise than the previous in-window gym visit, or if the same were true for out-of-window gym visits we would expect to see a non-linear relationship between past exercise and subsequent exercise. As mentioned above, the results in Appendix Table 41 do not show this pattern.

Another possible explanation for the finding that an incremental in-window gym visit generates less exercise in the future than an incremental out-of-window gym visit involves each incremental gym visit's intrinsic value, which we define as the value that an individual places on the gym visit in the absence of financial incentives. Consider a participant in one of the flexible experimental conditions. The financial incentives cause this participant to engage in both in-window and out-of-window gym visits on occasions when the participant would otherwise not visit the gym, meaning the gym visit's intrinsic value is slightly negative. In this case, the marginal in-window and out-of-window gym visits have quite similar intrinsic values, anywhere from zero to the negative of the amount of the financial incentive. Now consider a participant in one of the routine experimental conditions. The financial incentives cause this participant to engage in in-window gym visits with slightly negative intrinsic value, but there are no corresponding marginal out-of-window gym visits because out-of-window visits are not incentivized. Furthermore, financial incentives sometimes cause the participant to engage in in-window gym visits when the participant would have otherwise engaged in out-of-window gym visits. In these cases, the marginal out-of-window gym visits have slightly positive intrinsic value (between zero and the intrinsic value of the marginal in-window gym visit plus the amount of the financial incentive), while the marginal in-window gym visits have intrinsic values that are between the intrinsic value of the out-of-window gym visit and the intrinsic value of the out-of-window gym visit minus the amount of the financial incentive. Considering the flexible and routine experimental conditions together, the marginal in-window gym visits have lower intrinsic values than the marginal out-of-window gym visits, a fact that could explain our main finding if gym visits with higher intrinsic values lead to more subsequent gym visits.

A related possibility is that out-of-window gym visits are important for generating resilient habits. In-window gym visits may establish routines, and we indeed find that in-window gym visits during the incentive period are more effective than out-of-window visits during the incentive period at generating in-window visits after the incentive period. However, if an individual is unable to follow through on a routine gym visit 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 experimental conditions encouraged this behavior by incentivizing out-of-

study who were incentivized to badge into the gym (on March 25 and March 31, 2015). A newer gym audited on March 25 had a 79% badge-in rate and a 62% badge-out rate, while an older gym audited on March 31 had a 60% badge-in rate and 32% badge-out rate.

window exercise, and in the instrumental variables regressions we find that out-of-window gym visits during the incentive period lead to much more out-of-window exercise after the incentive period than do in-window gym visits during the incentive period, a pattern suggesting that the resilience to find an alternative time to exercise in the face of scheduling conflicts can be fostered by repetition of the behavior.

Finally, because our instrumental variables estimates are local average treatment effects for compliers, it is possible that our main result is driven by differences between the groups of compliers that are relevant for the in-window and out-of-window gym visit estimates. For example, the compliers whose out-of-window exercise behavior responds most strongly to the experimental conditions—and who are therefore most relevant for the estimate of the effect of out-of-window gym visits—may be the types of people who are most resilient in their ability to find alternative times to exercise and who are as a result most successful at forming exercise habits.

While we cannot disentangle the explanations having to do with the intrinsic value of marginal in-window and out-of-window gym visits, the importance of resilience in exercise habits, and possible differences between the groups of compliers relevant for the in-window and out-of-window gym visit estimates, all of these potential interpretations offer the same important lesson for 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, it is challenging to design interventions that take advantage of this insight.

Limitations

In spite of its scale and scope, our study has a number of important limitations. First, we are unable to monitor participants' exercise habits outside of their company gym. Although we attempted to measure all steps participants took by collecting data from wearable devices, only 5% of participants elected to provide this data. Second, we cannot observe participants' pre-intervention exercise patterns because the devices we used for tracking gym entrances and exits were installed (and participants' IDs were registered with those devices) at the time of our study's launch. Third, although participants reported no differences in their likihood of badging into the gym across experimental conditions (vs. failing to badge in; see "Validation of Data on Gym Usage"), the possibility that some gym visits were unobserved is an important limitation of our study. Fourth, our study was conducted at a single Fortune 500 company, and although it included 2,500 employees in a wide range of different job functions, this population is not representative of the U.S. workforce. Although we found no evidence of heterogeneous treatment effects by job type, the impact of inducing additional workouts at routine times might be different in organizations that structure work more or less flexibly. Our findings might also have differed if we had conducted our study at a gym that wasn't located at participants' place of work. Finally, in our study, all participants signed up for our four-week intervention with a "workout buddy". The effects of our intervention might have differed had participants signed up alone rather than with a partner.

5. CONCLUSION

This study examines the relative benefits of competing habit-formation interventions: incentivizing specific, routine behaviors versus incentivizing those behaviors flexibly. In a large

field experiment, we find that routine incentives, paid when a behavior is performed at a specific time, actually create less persistent habits than do flexible incentives, paid whenever a behavior is performed. It appears that this difference is driven by a form of crowding out. Specifically, while flexible incentives encouraged a general increase in exercise (without changing the amount of routine exercise), routine incentives encouraged only a small increase in exercise in a narrow time interval, often by simply shifting the time when exercise would have otherwise occurred. As a result, the average impact of incentivizing one additional routine workout was lower than the average impact of incentivizing one additional workout at any time, contrary to the prediction of a majority of leading social and cognitive psychologists surveyed at top U.S. universities.

Our study raises a number of important questions for future research. For instance, we examined whether a completely flexible incentive scheme was more effective at creating lasting habits than a routine incentive scheme. Of course, there may be a middle ground that is more effective than either of these extremes. For instance, an incentive scheme that pays participants for all workouts, but pays *more* for in-window workouts might help people build an exercise routine without discouraging participants who miss their workout window from exercising at another time. Future research exploring this possibility would be valuable. In addition, we defined routines at a daily (rather than weekly or monthly) interval and defined workout windows as two-hour periods. Relaxing or altering some of these definitions might have yielded different results, and exploring this would be an interesting direction for future research. Finally, a routine incentive may be more (or less) effective in a social context than in an individual context. A workout partner who is required to stay on the same schedule to earn incentives may provide extra support and accountability, helping participants stick to their routines, in addition to making workouts more enjoyable. This, in turn, may make routines stickier than they would be otherwise. However, a workout partner's failure to exercise may also license an individual to skip the gym, so a social routine could also be less sticky. Exploring this question in future research would be valuable.

An important issue to address is how cost-effectively our incentives increased exercise. We can conduct a rough calculation to determine how much it cost us, on average, to incentivize an additional workout for a given participant. The average participant in an incentivized experimental condition in our study visited the gym 29.9 total times during our 4-week intervention and the 40-weeks post-intervention follow-up period that we were able to observe. The average participant in our control condition visited the gym 22.8 times over that same, 44 week period. Our incentives, therefore, induced an estimated 7.1 additional workouts over the course of 44-weeks. Because we paid 2,376 participants in our incentive conditions $79,456 during the intervention period, the average incentive paid out per participant was $33.44 (=$79,456/2,376). This means that each additional workout our incentives induced cost approximately $4.71 (=$33.44/7.1). Whether this cost is outweighed by the benefits to cardiovascular health, mental health, and productivity associated with an extra visit to the gym is an important, open question (see Valero-Elizondo et al., 2016 for a discussion).

In conclusion, we present the results of an experiment exploring different strategies for promoting habit formation with important implications for policymakers interested in encouraging long-term behavior change. In spite of past reseach on habit formation suggesting that repeatedly rewarding valued behaviors under stable, cued conditions might promote more lasting habits than repeatedly rewarding valued behaviors flexibly, we find just the opposite. Exclusively incentivizing routines does not create more persistent habits than offering general incentives for desriable behaviors.

REFERENCES

Acland, D., & Levy, M. R. (2015). Naiveté, Projection Bias, and Habit Formation in Gym Attendance. *Management Science, 61*(1), 146-160. doi:doi:10.1287/mnsc.2014.2091

Alison Phillips, L., Leventhal, H., & Leventhal, E. A. (2013). Assessing theoretical predictors of long-term medication adherence: Patients' treatment-related beliefs, experiential feedback and habit development. *Psychology & Health, 28*(10), 1135-1151. doi:10.1080/08870446.2013.793798

Andersen, R. E. (1999). Exercise, an Active Lifestyle, and Obesity. *The Physician and Sportsmedicine, 27*(10), 41-50. doi:10.1080/00913847.1999.11439359

Becker, G. S., & Murphy, K. M. (1988). A Theory of Rational Addiction. *Journal of Political Economy, 96*(4), 675-700. doi:doi:10.1086/261558

Benartzi, S., Beshears, J., Milkman, K.L., Sunstein, C., Thaler, R.H., Shankar, M., Tucker, W., Congdon, W.J., and Galing S. (*forthcoming*). Should Governments Invest More in Nudging? *Psychological Science*.

Beshears, J., Choi, J. J., Laibson, D., & Madrian, B. C. (2013). Simplification and saving. *Journal of Economic Behavior & Organization, 95*, 130-145. doi:http://dx.doi.org/10.1016/j.jebo.2012.03.007

Brooks, T. L., Leventhal, H., Wolf, M. S., O'Conor, R., Morillo, J., Martynenko, M., . . . Federman, A. D. (2014). Strategies Used by Older Adults with Asthma for Adherence to Inhaled Corticosteroids. *Journal of General Internal Medicine, 29*(11), 1506-1512. doi:10.1007/s11606-014-2940-8

Brownell, K. D. (1995). Exercise in the treatment of obesity. *Eating disorders and obesity: A comprehensive handbook*, 473-478.

"Exercise or Physical Activity." *Centers for Disease Control and Prevention*. Centers for Disease Control and Prevention, 07 Oct. 2016. Web. 13 Feb. 2017.

Charness, G., & Gneezy, U. (2009). Incentives to Exercise. *Econometrica, 77*(3), 909-931. doi:10.3982/ECTA7416

Cochran, W., & Tesser, A. (1996). The "what the hell" effect: Some effects of goal proximity and goal framing on performance. *Striving and feeling: Interactions among goals, affect, and self-regulation*, 99-120.

Ding, D., Lawson, K. D., Kolbe-Alexander, T. L., Finkelstein, E. A., Katzmarzyk, P. T., van Mechelen, W., & Pratt, M. (2016). The economic burden of physical inactivity: a global analysis of major non-communicable diseases. *The Lancet, 388*(10051), 1311-1324. doi:http://dx.doi.org/10.1016/S0140-6736(16)30383-X

Flegal, K. M., Carroll, M. D., Ogden, C. L., & Curtin, L. R. (2010). Prevalence and trends in obesity among us adults, 1999-2008. *JAMA, 303*(3), 235-241. doi:10.1001/jama.2009.2014

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., . . . Grantham-McGregor, S. (2014). Labor market returns to an early childhood stimulation intervention in Jamaica. *Science, 344*(6187), 998-1001. doi:10.1126/science.1251178

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist, 54*(7), 493-503. doi:10.1037/0003-066X.54.7.493

Gruber, Jonathan. "Tobacco at the Crossroads: The Past and Future of Smoking Regulation in the United States." *The Journal of Economic Perspectives*, 15.2 (2001): 193–212.

Gruber, Jonathan, and Botond Köszegi. "Is addiction "rational"? Theory and evidence." *The Quarterly Journal of Economics* 116.4 (2001): 1261-1303.

Gruber, Jonathan, and Botond Kőszegi. "Tax incidence when individuals are time-inconsistent: the case of cigarette excise taxes." *Journal of Public Economics* 88.9 (2004): 1959-1987.

Johnson, E. J., & Goldstein, D. (2003). Do Defaults Save Lives? *Science, 302*(5649), 1338-1339. doi:10.1126/science.1091721

Judah, G., Gardner, B., & Aunger, R. (2013). Forming a flossing habit: An exploratory study of the psychological determinants of habit formation. *British Journal of Health Psychology, 18*(2), 338-353. doi:10.1111/j.2044-8287.2012.02086.x

Kayman, S., Bruvold, W., & Stern, J. S. (1990). Maintenance and relapse after weight loss in women: behavioral aspects. *The American Journal of Clinical Nutrition, 52*(5), 800-807.

Khera, A. V., Emdin, C. A., Drake, I., Natarajan, P., Bick, A. G., Cook, N. R., . . . Kathiresan, S. (2016). Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *New England Journal of Medicine, 375*(24), 2349-2358. doi:doi:10.1056/NEJMoa1605086

Kuh, G., Kinzie, J., Buckley, J., Bridges, B., & Hayek, J. (2006). What matters to student success: A review of the literature. Washington, DC: National Postsecondary Education Cooperative. *National Center for Education Statistics*.

Ladenvall, P., Persson, C. U., Mandalenakis, Z., Wilhelmsen, L., Grimby, G., Svärdsudd, K., & Hansson, P.-O. (2016). Low aerobic capacity in middle-aged men associated with increased mortality rates during 45 years of follow-up. *European Journal of Preventive Cardiology, 23*(14), 1557-1564. doi:10.1177/2047487316655466

Larrick, R. P., & Soll, J. B. (2008). The MPG Illusion. *Science, 320*(5883), 1593-1594. doi:10.1126/science.1154983

Lee, C. D., Blair, S. N., & Jackson, A. S. (1999). Cardiorespiratory fitness, body composition, and all-cause and cardiovascular disease mortality in men. *The American Journal of Clinical Nutrition, 69*(3), 373-380.

Loewenstein, G., Price, J., & Volpp, K. (2016). Habit formation in children: Evidence from incentives for healthy eating. *Journal of Health Economics, 45*, 47-54. doi:http://dx.doi.org/10.1016/j.jhealeco.2015.11.004

Madrian, B. C., & Shea, D. F. (2001). The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior*. *The Quarterly Journal of Economics, 116*(4), 1149-1187. doi:10.1162/003355301753265543

Milkman, K. L., Beshears, J., Choi, J. J., Laibson, D., & Madrian, B. C. (2011). Using implementation intentions prompts to enhance influenza vaccination rates. *Proceedings of the National Academy of Sciences, 108*(26), 10415-10420. doi:10.1073/pnas.1103170108

Milkman, K. L., Minson, J. A., & Volpp, K. G. M. (2014). Holding the Hunger Games Hostage at the Gym: An Evaluation of Temptation Bundling. *Management Science, 60*(2), 283-299. doi:doi:10.1287/mnsc.2013.1784

Mokdad, A. H., Marks, J. S., Stroup, D. F., & Gerberding, J. L. (2004). Actual causes of death in the united states, 2000. *JAMA, 291*(10), 1238-1245. doi:10.1001/jama.291.10.1238

Patel, M. S., Asch, D. A., Rosin, R., & et al. (2016). Framing financial incentives to increase physical activity among overweight and obese adults: A randomized, controlled trial. *Annals of Internal Medicine, 164*(6), 385-394. doi:10.7326/M15-1635

Polivy, J. (1976). Perception of calories and regulation of intake in restrained and unrestrained subjects. *Addictive Behaviors, 1*(3), 237-243. doi:http://dx.doi.org/10.1016/0306-4603(76)90016-2

Rogers, T., Milkman, K. L., John, L. K., & Norton, M. I. (2015). Beyond good intentions: Prompting people to make plans improves follow-through on important tasks. *Behavioral Science & Policy, 1*(2), 33-41.

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. *American Economic Journal: Applied Economics, 7*(3), 51-84. doi:doi: 10.1257/app.20130327

Schroeder , S. A. (2007). We Can Do Better — Improving the Health of the American People. *New England Journal of Medicine, 357*(12), 1221-1228. doi:doi:10.1056/NEJMsa073350

Sen, A. P., Sewell, T. B., Riley, E. B., Stearman, B., Bellamy, S. L., Hu, M. F., . . . Volpp, K. G. (2014). Financial Incentives for Home-Based Health Monitoring: A Randomized Controlled Trial. *Journal of General Internal Medicine, 29*(5), 770-777. doi:10.1007/s11606-014-2778-0

Sharif, M. A., & Shu, S. B. (in press). The Benefits of Emergency Reserves: Greater Preference and Persistence for Goals having Slack with a Cost. *Journal of Marketing Research, 0*(ja), null. doi:10.1509/jmr.15.0231

Soman, D., & Cheema, A. (2004). When Goals Are Counterproductive: The Effects of Violation of a Behavioral Goal on Subsequent Performance. *Journal of Consumer Research, 31*(1), 52-62. doi:10.1086/383423

Staats, B. R., Dai, H., Hofmann, D., & Milkman, K. L. (forthcoming). Motivating Process Compliance Through Individual Electronic Monitoring: An Empirical Examination of Hand Hygiene in Healthcare. *Management Science, 0*(0), null. doi:10.1287/mnsc.2015.2400

Tappe, K., Tarves, E., Oltarzewski, J., & Frum, D. (2013). Habit Formation among Regular Exercisers at Fitness Centers: An Exploratory Study. *Journal of Physical Activity and Health, 10*(4), 607-613. doi:doi:10.1123/jpah.10.4.607

Team, S. a. B. S. (2016). *2016 Annual Report*. Retrieved from https://sbst.gov/download/2016%20SBST%20Annual%20Report.pdf.

Thaler, R. H., & Benartzi, S. (2004). Save more tomorrow™: Using behavioral economics to increase employee saving. *Journal of Political Economy, 112*(S1), S164-S187.

Thaler, R. H., & Sunstein, C. (2008). Nudge: Improving decisions about health, wealth, and happiness: New Haven, CT: Yale University Press.

U.S. News & World Report (2016). U.S. News & World Report, L.P. https://www.usnews.com/. Accessed May 20, 2017.

Valero-Elizondo, et al (2016). Economic Impact of Moderate-Vigorous Physical Activity Among Those With and Without Established Cardiovascular Disease: 2012 Medical Expenditure Panel Survey. *Journal of the American Heart Association*, 5.9.

Verplanken, B., & Wood, W. (2006). Interventions to Break and Create Consumer Habits. *Journal of Public Policy & Marketing, 25*(1), 90-103. doi:10.1509/jppm.25.1.90

Wood, W., & Neal, D. T. (2016). Healthy through habit: Interventions for initiating & maintaining health behavior change. *Behavioral Science & Policy, 2*(1), 71-83.

TABLES AND FIGURES

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

	Total	Control	Flexible			Routine		
			Overall	*$3*	*$7*	*Overall*	*$3*	*$7*
Age	31.42	31.68	31.52	31.90	31.15	31.27	31.69	30.83
	(6.59)	(6.10)	(6.66)	(7.04)	(6.26)	(6.57)	(6.90)	(6.20)
Proportion that responded	69%	75%	71%	70%	72%	67%	68%	66%
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 (Pre-Intervention)	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	24.78	24.36	24.79	24.82	24.77	24.82	24.76	24.89
	(4.33)	(4.09)	(4.55)	(4.72)	(4.39)	(4.12)	(4.22)	(4.02)
Proportion that responded	67%	75%	69%	68%	70%	65%	67%	63%
Proportion of Males (%)	54.91	54.69	53.14	52.46	53.83	56.72	57.75	55.68
Proportion that responded	95%	97%	95%	95%	94%	94%	94%	94%
Job Function								
Tech (%)	60.65	64.65	59.69	59.23	60.14	61.17	62.47	59.79
Global Business Organization (%)	20.72	13.13	21.39	24.22	18.66	20.95	21.98	19.84
General & Administrative (%)	18.64	22.22	18.92	16.55	21.20	17.88	15.56	20.37
Proportion that responded	69%	75%	71%	70%	73%	66%	68%	64%
Job Level								
L1 (%)	0.94	1.10	0.89	0.78	0.99	0.97	0.54	1.42
L2 (%)	8.05	6.59	8.88	6.53	11.11	7.32	7.26	7.39
L3 (%)	26.64	24.18	25.63	24.80	26.42	28.04	26.61	29.55
L4 (%)	31.94	24.18	31.85	30.29	33.33	33.01	32.53	33.52
L5 (%)	19.90	27.47	20.05	22.72	17.53	18.78	19.35	18.18
L6 (%)	7.92	10.99	8.50	9.40	7.65	6.91	8.06	5.68
L7 (%)	2.62	1.10	2.41	2.87	1.98	3.04	3.23	2.84
L8+ (%)	0.81	1.10	0.89	1.57	0.25	0.69	0.54	0.85
Intern (%)	0.56	0.00	0.51	0.78	0.25	0.69	0.81	0.57
other (%)	0.62	3.30	0.38	0.26	0.49	0.55	1.08	0.00
Proportion that responded	64%	69%	66%	64%	68%	61%	63%	60%
Ethnicity								
White (%)	49.17	51.67	49.29	48.87	49.71	48.76	49.62	47.90
Black (%)	2.79	3.33	2.66	2.44	2.88	2.86	3.24	2.48
Asian (%)	35.88	35.00	34.47	35.90	33.01	37.40	37.02	37.79
Hispanic (%)	4.73	3.33	5.98	5.64	6.33	3.63	3.63	3.63
Mixed or Other (%)	7.43	6.67	7.60	7.14	8.06	7.35	6.49	8.21
Proportion that responded	89%	91%	88%	89%	88%	89%	88%	89%
Sample Size	2,508	132	1,194	600	594	1,182	594	588

Standard errors robust in parentheses.

Note: This table summarizes key employee characteristics based on responses to questions included in our registration survey, which participants had the option to skip. Since responding to these questions was voluntary, we report the proportion of participants who responded to each question.

Table 2. Summary Statistics Describing Participants' Levels of Engagement with Our Experiment Overall and By Condition

	Total	Control	Flexible			Routine		
			Total	*$3*	*$7*	*Total*	*$3*	*$7*
Email Reminders								
Initial Opt-In	100%	100%	100%	100%	100%	100%	100%	100%
Post-Randomization Opt-In	99%	99%	99%	98%	99%	98%	98%	98%
Text Reminders								
Initial Opt-In	42%	48%	40%	40%	41%	43%	44%	43%
Post-Randomization Opt-In	36%	41%	34%	32%	35%	37%	37%	37%
Workout Windows (Starting Time)								
Morning (3:00AM to 8:45AM)	22%	18%	22%	23%	22%	22%	21%	23%
Midday (9:00AM to 2:45PM)	29%	34%	30%	31%	30%	27%	28%	26%
Afternoon (3:00PM to 8:45PM)	48%	45%	46%	44%	47%	50%	50%	49%
Evening (9:00PM to 2:45AM)	1%	1%	2%	2%	1%	1%	1%	2%
Perfect Overlap with Partner's Window	32%	35%	30%	30%	30%	33%	33%	32%
Enrolled in AchieveMint	26%	28%	26%	27%	25%	25%	26%	25%
Total Sample Size	2,508	132	1,194	600	594	1,182	594	588

Note: This table reports on summary statistics that offer insight into our participants' level of engagement with our experiment. Responses were required to these questions. Due to a glitch on our survey, a single individual in the Flexible-$3 Incentive group did not select a workout window. We control for this single individual in our regressions, effectively excluding this individual from our analysis. These summary statistics are all as of the end of the intervention period, while the control variables used in regression analysis were measured as of the date of randomization (to ensure no endogeneity issues in our control variables). These summary statistics as of the end of the intervention period are more interesting to display (for example, the percentage of participants receiving email reminders is 100% as of the date of randomization because of the study design), but the summary statistics as of the date of randomization are very similar.

Table 3. Panel A. Regressions Predicting Participants' Weekly Workouts during the Intervention Period

This table reports a series of ordinary least squares regressions predicting a study participant's weekly average number of (a) overall workouts, (b) workouts initiated during their workout window, and (c) workouts initiated outside of their workout window during our four week intervention period. Columns (2) and (3) are the first stage IV regressions described in Equation 1. 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 versus $7) and the flexibility of the workout schedule (flexible versus routine). We report pair-wise Wald Tests to assess whether or not all paired regression coefficients reported differ significantly from each other.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Total Qualifying Workouts	Total Qualifying In-Window Workouts	Total Qualifying Out-of-Window Workouts	Total Qualifying Workouts	Total Qualifying In-Window Workouts	Total Qualifying Out-of-Window Workouts	Total Qualifying Workouts	Total Qualifying In-Window Workouts	Total Qualifying Out-of-Window Workouts
Flexible Payment $3	0.59***	0.32***	0.26**						
	(0.14)	(0.09)	(0.09)						
Flexible Payment $7	0.90***	0.43***	0.46***						
	(0.14)	(0.10)	(0.09)						
Routine Payment $3	0.40**	0.58***	-0.18*						
	(0.14)	(0.10)	(0.08)						
Routine Payment $7	0.69***	0.97***	-0.27**						
	(0.14)	(0.10)	(0.08)						
$3 Interventions				0.49***	0.45***	0.04			
				(0.13)	(0.09)	(0.08)			
$7 Interventions				0.80***	0.70***	0.10			
				(0.13)	(0.09)	(0.09)			
Flexible Interventions							0.74***	0.38***	0.36***
							(0.13)	(0.09)	(0.09)
Routine Interventions							0.55***	0.77***	-0.22**
							(0.13)	(0.09)	(0.08)
Mean Values of Control Group	1.11	0.59	0.52	1.11	0.59	0.52	1.11	0.59	0.52
Observations	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032
R-squared	0.07	0.07	0.11	0.07	0.05	0.03	0.07	0.06	0.10

Table 3. Panel A. Regressions Predicting Participants' Weekly Workouts during the Intervention Period (continued)

	(1) Total Qualifying Workouts	(2) Total Qualifying In-Window Workouts	(3) Total Qualifying Out-of-Window Workouts	(4) Total Qualifying Workouts	(5) Total Qualifying In-Window Workouts	(6) Total Qualifying Out-of-Window Workouts	(7) Total Qualifying Workouts	(8) Total Qualifying In-Window Workouts	(9) Total Qualifying Out-of-Window Workouts
Wald Test ($3 Flexible-$7 Flexible)									
Difference in Coefficients	-0.31***	-0.11	-0.20**						
	(0.09)	(0.07)	(0.06)						
Wald Test ($3 Flexible-$3 Routine)									
Difference in Coefficients	0.19*	-0.25**	0.44***						
	(0.09)	(0.08)	(0.05)						
Wald Test ($3 Flexible-$7 Routine)									
Difference in Coefficients	-0.11	-0.64***	0.53***						
	(0.09)	(0.08)	(0.05)						
Wald Test ($7 Flexible-$3 Routine)									
Difference in Coefficients	0.50***	-0.14+	0.64***						
	(0.09)	(0.08)	(0.05)						
Wald Test ($7 Flexible-$7 Routine)									
Difference in Coefficients	0.20*	-0.53***	0.73***						
	(0.09)	(0.08)	(0.05)						
Wald Test ($3 Routine-$7 Routine)									
Difference in Coefficients	-0.30***	-0.39***	0.09**						
	(0.09)	(0.09)	(0.03)						
Wald Test ($3-$7)									
Difference in Coefficients				-0.30***	-0.25***	-0.05			
				(0.06)	(0.06)	(0.04)			
Wald Test (Flexible-Routine)									
Difference in Coefficients							0.19**	-0.39***	0.59***
							(0.06)	(0.06)	(0.04)

Note: Standard errors robust to heteroscedasticity and clustered by workout work-out buddy pair in parentheses. +p<0.10, *p<0.05, **p<0.01, ***p<0.001
All regressions contain indicators controlling for our three stratification dates, perfect overlap with partner's workout window (y/n), and above median self-reported workouts per week (y/n). Additionally, we include 12 interaction effects between our stratification dates, perfect overlap with partner's workout window (y/n), and above media self-reported workouts per week (y/n).

Table 3. Panel B. Regressions Predicting Participants' Likelihood of Working out Each Week during the Intervention Period
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 their workout window, and (c) workout initiated outside of their workout window during our four week intervention period. In each column, we report the mean number 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 versus $7) and the flexibility of the workout schedule (flexible versus routine). We report pair-wise Wald Tests to assess whether or not all paired regression coefficients reported differ significantly from each other.

	(1) Any Qualifying Workouts? (Y/N)	(2) Any Qualifying In-Window Workouts? (Y/N)	(3) Any Qualifying Out-of-Window Workouts? (Y/N)	(4) Any Qualifying Workouts? (Y/N)	(5) Any Qualifying In-Window Workouts? (Y/N)	(6) Any Qualifying Out-of-Window Workouts? (Y/N)	(7) Any Qualifying Workouts? (Y/N)	(8) Any Qualifying In-Window Workouts? (Y/N)	(9) Any Qualifying Out-of-Window Workouts? (Y/N)
Flexible Payment $3	0.16***	0.14***	0.13***						
	(0.04)	(0.04)	(0.04)						
Flexible Payment $7	0.23***	0.17***	0.19***						
	(0.04)	(0.04)	(0.04)						
Routine Payment $3	0.13**	0.21***	-0.07+						
	(0.04)	(0.04)	(0.04)						
Routine Payment $7	0.17***	0.30***	-0.12***						
	(0.04)	(0.04)	(0.03)						
$3 Interventions				0.14***	0.17***	0.03			
				(0.04)	(0.03)	(0.03)			
$7 Interventions				0.20***	0.24***	0.03			
				(0.04)	(0.03)	(0.03)			
Flexible Interventions							0.19***	0.15***	0.16***
							(0.04)	(0.03)	(0.03)
Routine Interventions							0.15***	0.25***	-0.09**
							(0.04)	(0.03)	(0.03)
Mean Values of Control Group	0.50	0.31	0.30	0.50	0.31	0.30	0.50	0.31	0.30
Observations	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032
R-squared	0.05	0.06	0.10	0.05	0.05	0.03	0.05	0.05	0.09

Table 3. Panel B. Regressions Predicting Participants' Likelihood of Working out Each Week during the Intervention Period (continued)

	(1) Any Qualifying Workouts? (Y/N)	(2) Any Qualifying In-Window Workouts? (Y/N)	(3) Any Qualifying Out-of-Window Workouts? (Y/N)	(4) Any Qualifying Workouts? (Y/N)	(5) Any Qualifying In-Window Workouts? (Y/N)	(6) Any Qualifying Out-of-Window Workouts? (Y/N)	(7) Any Qualifying Workouts? (Y/N)	(8) Any Qualifying In-Window Workouts? (Y/N)	(9) Any Qualifying Out-of-Window Workouts? (Y/N)
Wald Test ($3 Flexible-$7 Flexible)									
Difference in Coefficients	-0.07**	-0.04	-0.06*						
	(0.02)	(0.02)	(0.02)						
Wald Test ($3 Flexible-$3 Routine)									
Difference in Coefficients	0.03	-0.07**	0.19***						
	(0.02)	(0.03)	(0.02)						
Wald Test ($3 Flexible-$7 Routine)									
Difference in Coefficients	-0.02	-0.16***	0.25***						
	(0.02)	(0.02)	(0.02)						
Wald Test ($7 Flexible-$3 Routine)									
Difference in Coefficients	0.10***	-0.03	0.26***						
	(0.02)	(0.03)	(0.02)						
Wald Test ($7 Flexible-$7 Routine)									
Difference in Coefficients	0.05*	-0.13***	0.31***						
	(0.02)	(0.02)	(0.02)						
Wald Test ($3 Routine-$7 Routine)									
Difference in Coefficients	-0.04+	-0.09***	0.05**						
	(0.02)	(0.03)	(0.02)						
Wald Test ($3-$7)									
Difference in Coefficients				-0.06***	-0.07***	-0.004			
				(0.02)	(0.02)	(0.02)			
Wald Test (Flexible-Routine)									
Difference in Coefficients							0.04*	-0.10***	0.25***
							(0.02)	(0.02)	(0.01)

Note: Standard errors robust to heteroscedasticity and clustered by workout work-out buddy pair in parentheses. +p<0.10, *p<0.05, **p<0.01, ***p<0.001. All regressions contain indicators controlling for our three stratification dates, perfect overlap with partner's workout window (y/n), and above median self-reported workouts per week (y/n). Additionally, we include 12 interaction effects between our stratification dates, perfect overlap with partner's workout window (y/n), and above media self-reported workouts per week (y/n).

Table 4. Panel A. Regressions Predicting Participants' Weekly Workouts during the 4 Weeks Post-Intervention

This table reports a series of ordinary least squares regressions predicting a study participant's weekly average number of (a) overall workouts, (b) workouts initiated during their workout window, and (c) workouts initiated outside of their workout window during the four weeks following our intervention period. Columns (2) and (3) are the first stage IV regressions described in Equation 1. 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 versus $7) and the flexibility of the workout schedule (flexible versus routine). We report pair-wise Wald Tests to assess whether or not all paired regression coefficients reported differ significantly from each other.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Total Qualifying Workouts	Total Qualifying In-Window Workouts	Total Qualifying Out-of-Window Workouts	Total Qualifying Workouts	Total Qualifying In-Window Workouts	Total Qualifying Out-of-Window Workouts	Total Qualifying Workouts	Total Qualifying In-Window Workouts	Total Qualifying Out-of-Window Workouts
Flexible Payment $3	0.21+	0.07	0.14*						
	(0.11)	(0.08)	(0.06)						
Flexible Payment $7	0.28*	0.09	0.19**						
	(0.11)	(0.08)	(0.06)						
Routine Payment $3	0.12	0.07	0.05						
	(0.11)	(0.08)	(0.06)						
Routine Payment $7	0.18	0.15+	0.02						
	(0.11)	(0.08)	(0.06)						
$3 Interventions				0.17	0.07	0.10			
				(0.11)	(0.07)	(0.06)			
$7 Interventions				0.23*	0.12	0.11+			
				(0.11)	(0.07)	(0.06)			
Flexible Interventions							0.25*	0.08	0.17**
							(0.11)	(0.07)	(0.06)
Routine Interventions							0.15	0.11	0.04
							(0.11)	(0.07)	(0.06)
Mean Values of Control Group	0.76	0.42	0.34	0.76	0.42	0.34	0.76	0.42	0.34
Observations	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032
R-squared	0.08	0.05	0.04	0.08	0.05	0.04	0.08	0.05	0.04

Table 4. Panel A. Regressions Predicting Participants' Weekly Workouts during the 4 Weeks Post-Intervention (continued)

	(1) Total Qualifying Workouts	(2) Total Qualifying In-Window Workouts	(3) Total Qualifying Out-of-Window Workouts	(4) Total Qualifying Workouts	(5) Total Qualifying In-Window Workouts	(6) Total Qualifying Out-of-Window Workouts	(7) Total Qualifying Workouts	(8) Total Qualifying In-Window Workouts	(9) Total Qualifying Out-of-Window Workouts
Wald Test ($3 Flexible-$7 Flexible)									
Difference in Coefficients	-0.07	-0.02	-0.05						
	(0.07)	(0.05)	(0.05)						
Wald Test ($3 Flexible-$3 Routine)									
Difference in Coefficients	0.09	0.002	0.09*						
	(0.07)	(0.05)	(0.04)						
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	0.14**						
	(0.07)	(0.05)	(0.05)						
Wald Test ($7 Flexible-$7 Routine)									
Difference in Coefficients	0.10	-0.06	0.17***						
	(0.07)	(0.06)	(0.04)						
Wald Test ($3 Routine-$7 Routine)									
Difference in Coefficients	-0.06	-0.08	0.03						
	(0.07)	(0.06)	(0.04)						
Wald Test ($3-$7)									
Difference in Coefficients				-0.06	-0.05	-0.01			
				(0.05)	(0.04)	(0.03)			
Wald Test (Flexible-Routine)									
Difference in Coefficients							0.10*	-0.03	0.13***
							(0.05)	(0.04)	(0.03)

Note: Standard errors robust to heteroscedasticity and clustered by workout work-out buddy pair in parentheses. +p<0.10, *p<0.05, **p<0.01, ***p<0.001
All regressions contain indicators controlling for our three stratification dates, perfect overlap with partner's workout window (y/n), and above median self-reported workouts per week (y/n). Additionally, we include 12 interaction effects between our stratification dates, perfect overlap with partner's workout window (y/n), and above media self-reported workouts per week (y/n).

Table 4. Panel B. Regressions Predicting Participants' Likelihood of Working out Each Week during the 4 Weeks Post-Intervention

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 their workout window, and (c) workout initiated outside of their workout window during the four weeks following our intervention period. In each column, we report the mean number 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 versus $7) and the flexibility of the workout schedule (flexible versus routine). We report pair-wise Wald Tests to assess whether or not all paired regression coefficients reported differ significantly from each other.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Any Qualifying Workouts? (Y/N)	Any Qualifying In-Window Workouts? (Y/N)	Any Qualifying Out-of-Window Workouts? (Y/N)	Any Qualifying Workouts? (Y/N)	Any Qualifying In-Window Workouts? (Y/N)	Any Qualifying Out-of-Window Workouts? (Y/N)	Any Qualifying Workouts? (Y/N)	Any Qualifying In-Window Workouts? (Y/N)	Any Qualifying 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	0.02						
	(0.04)	(0.03)	(0.03)						
Routine Payment $7	0.07+	0.06*	0.02						
	(0.04)	(0.03)	(0.03)						
$3 Interventions				0.07*	0.04	0.04			
				(0.03)	(0.03)	(0.03)			
$7 Interventions				0.10**	0.07*	0.06+			
				(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+	0.02
							(0.03)	(0.03)	(0.03)
Mean Values of Control Group	0.34	0.22	0.21	0.34	0.22	0.21	0.34	0.22	0.21
Observations	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032
R-squared	0.07	0.05	0.05	0.07	0.05	0.04	0.07	0.05	0.05

Table 4. Panel B. Regressions Predicting Participants' Likelihood of Working out Each Week during the 4 Weeks Post-Intervention (continued)

	(1) Any Qualifying Workouts? (Y/N)	(2) Any Qualifying In-Window Workouts? (Y/N)	(3) Any Qualifying Out-of-Window Workouts? (Y/N)	(4) Any Qualifying Workouts? (Y/N)	(5) Any Qualifying In-Window Workouts? (Y/N)	(6) Any Qualifying Out-of-Window Workouts? (Y/N)	(7) Any Qualifying Workouts? (Y/N)	(8) Any Qualifying In-Window Workouts? (Y/N)	(9) Any Qualifying Out-of-Window Workouts? (Y/N)
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*	-0.004	0.05**						
	(0.02)	(0.02)	(0.02)						
Wald Test ($3 Flexible-$7 Routine)									
Difference in Coefficients	0.03	-0.03	0.05**						
	(0.02)	(0.02)	(0.02)						
Wald Test ($7 Flexible-$3 Routine)									
Difference in Coefficients	0.08***	0.03	0.08***						
	(0.02)	(0.02)	(0.02)						
Wald Test ($7 Flexible-$7 Routine)									
Difference in Coefficients	0.07**	0.01	0.08***						
	(0.02)	(0.02)	(0.02)						
Wald Test ($3 Routine-$7 Routine)									
Difference in Coefficients	-0.02	-0.02	0.002						
	(0.02)	(0.02)	(0.02)						
Wald Test ($3-$7)									
Difference in Coefficients				-0.03	-0.03+	-0.01			
				(0.02)	(0.02)	(0.01)			
Wald Test (Flexible-Routine)									
Difference in Coefficients							0.06***	0.001	0.07***
							(0.02)	(0.02)	(0.01)

Note: Standard errors robust to heteroscedasticity and clustered by workout work-out buddy pair in parentheses. +p<0.10, *p<0.05, **p<0.01, ***p<0.001
All regressions contain indicators controlling for our three stratification dates, perfect overlap with partner's workout window (y/n), and above median self-reported workouts per week (y/n). Additionally, we include 12 interaction effects between our stratification dates, perfect overlap with partner's workout window (y/n), and above media self-reported workouts per week (y/n).

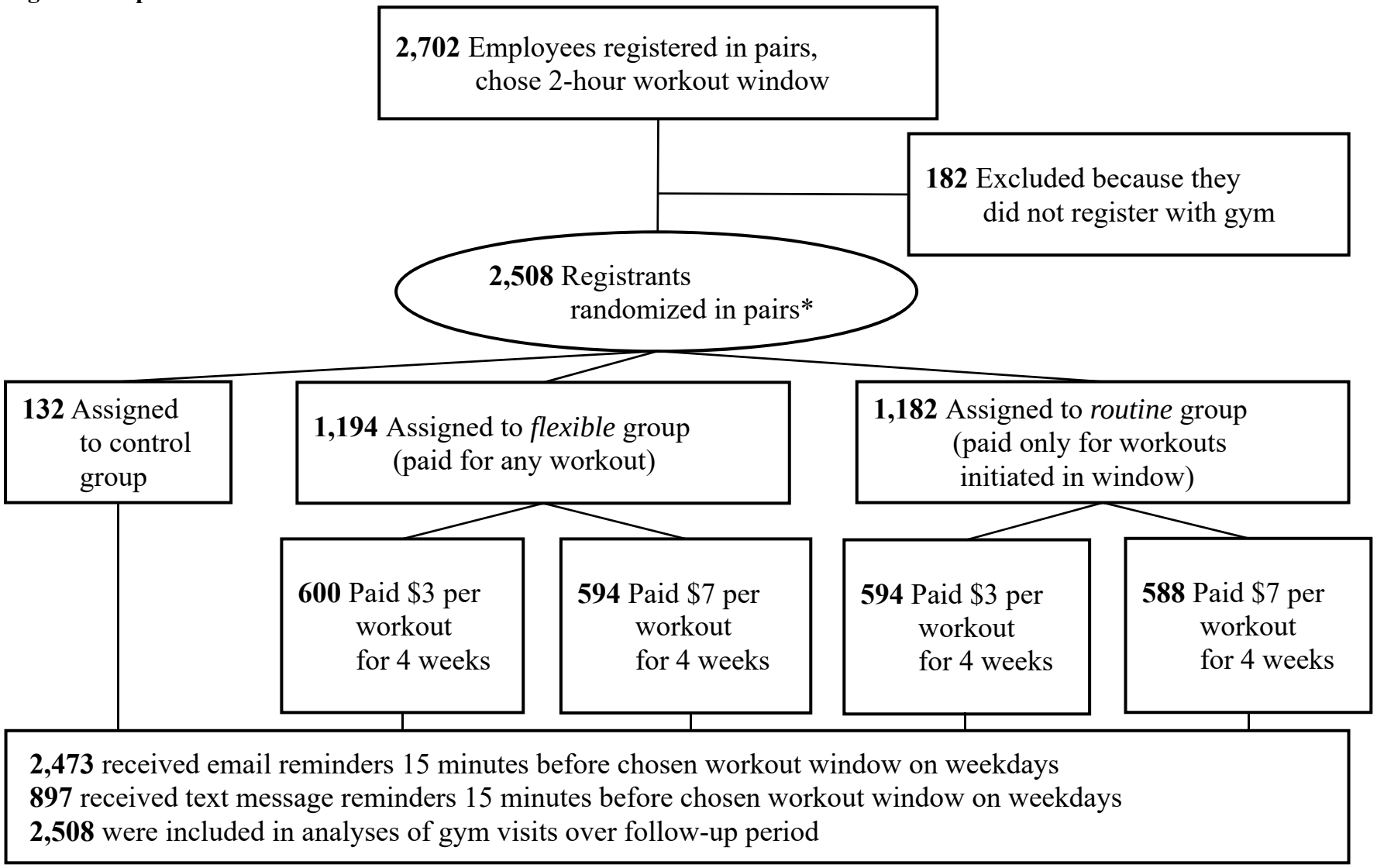
Table 5. IV Regressions, 4 Weeks Post-Intervention

This table reports on a series of second stage regressions from our IV regressions described in Equation 2 predicting a study participant's weekly average number of (a) overall workouts, (b) workouts initiated during their workout window, and (c) workouts initiated outside of their workout window during the four week period following our intervention as well as a participant's a study participant's weekly likelihood of completing a (a) workout anytime, (b) workout initiated during their workout window, and (c) workout initiated outside of their workout window. The key predictors are the total number of weekly in-window and out-of-window workouts completed during the intervention period of our study. We instrument for in-window and out-of-window workouts during our intervention period using treatment status. We report pair-wise Wald Tests to assess whether or not our estimated regression coefficients are significantly different from each other. Additionally, we test for underidentification and weak instruments, and the Kleibergen-Paap Wald rk F statistic associated with the first stage of our IV estimation is 13.87.

	Weekly Total Workouts 4-Weeks Post-Intervention		**Weekly In-Window Workouts 4-Weeks Post-Intervention**		**Weekly Out-of-Window Workouts 4-Weeks Post-Intervention**	
	(1) Total Qualifying Workouts	(2) Any Qualifying Workout? (Y/N)	(3) Total Qualifying Workouts	(4) Any Qualifying Workout? (Y/N)	(5) Total Qualifying Workouts	(6) Any Qualifying Workout? (Y/N)
Weekly In-Window Workouts	0.25*	0.11***	0.18*	0.09**	0.06	0.05
	(0.10)	(0.03)	(0.08)	(0.03)	(0.06)	(0.03)
Weekly Out-of-Window Workouts	0.33***	0.17***	0.06	0.06*	0.26***	0.15***
	(0.09)	(0.03)	(0.07)	(0.03)	(0.06)	(0.03)
Observations	10,032	10,032	10,032	10,032	10,032	10,032
R-squared	0.29	0.26	0.22	0.20	0.20	0.19
Wald Test						
Difference in Coefficients	-0.08	-0.06*	0.12+	0.02	-0.20***	-0.10***
	(0.08)	(0.03)	(0.07)	(0.02)	(0.05)	(0.02)

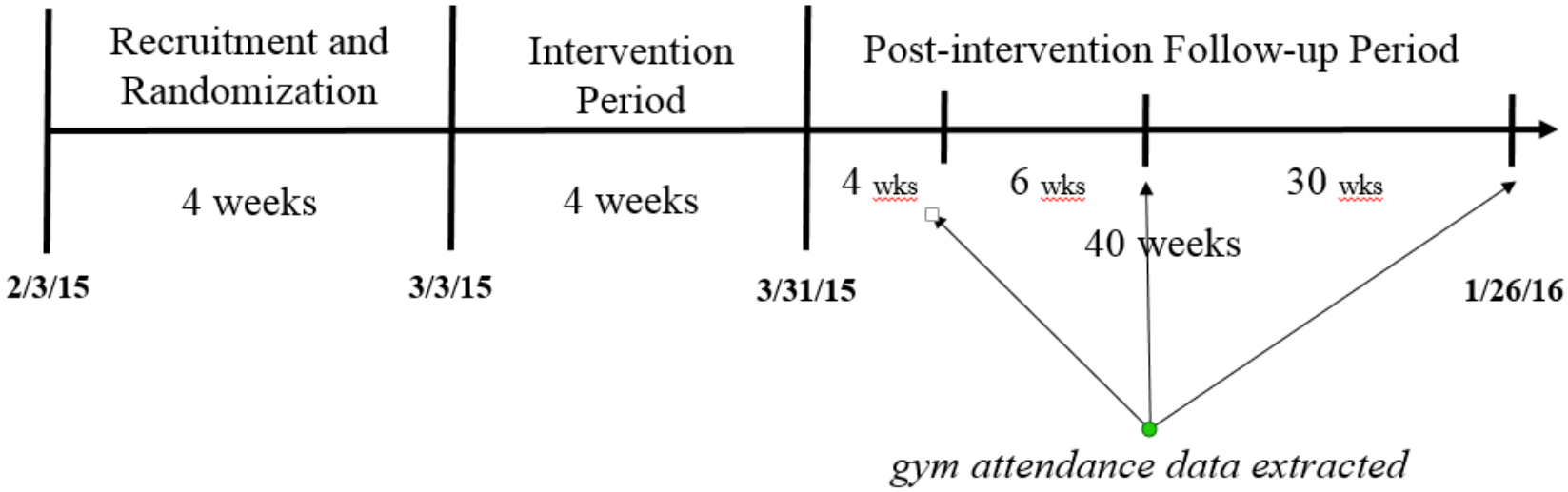
Note: Standard errors robust to heteroscedasticity and clustered by workout work-out buddy pair in parentheses. +p<0.10, *p<0.05, **p<0.01, ***p<0.001
All regressions contain indicators controlling for our three stratification dates, perfect overlap with partner's workout window (y/n), and above median self-reported workouts per week (y/n). Additionally, we include 12 interaction effects between our stratification dates, perfect overlap with partner's workout window (y/n), and above media self-reported workouts per week (y/n)..

Figure 1. Experimental Flow Chart



* Twelve registrants (1 in the *control group*, 3 in the *flexible $3* group, 1 in the *flexible $7* group, 6 in the *routine $3* group, and 1 in the *routine $7* group) were randomized but later opted-out of email notifications.

Figure 2. Study Timeline[14]



[14] All days up to and including 3/3/15 are considered pre-intervention

Figure 3. Panel A. Average Number of Workouts for Four Weeks During Our Intervention by Condition
This figure shows the average number workouts in a given week overall, in-window and out-of-window during our 4-week intervention period by experimental condition.

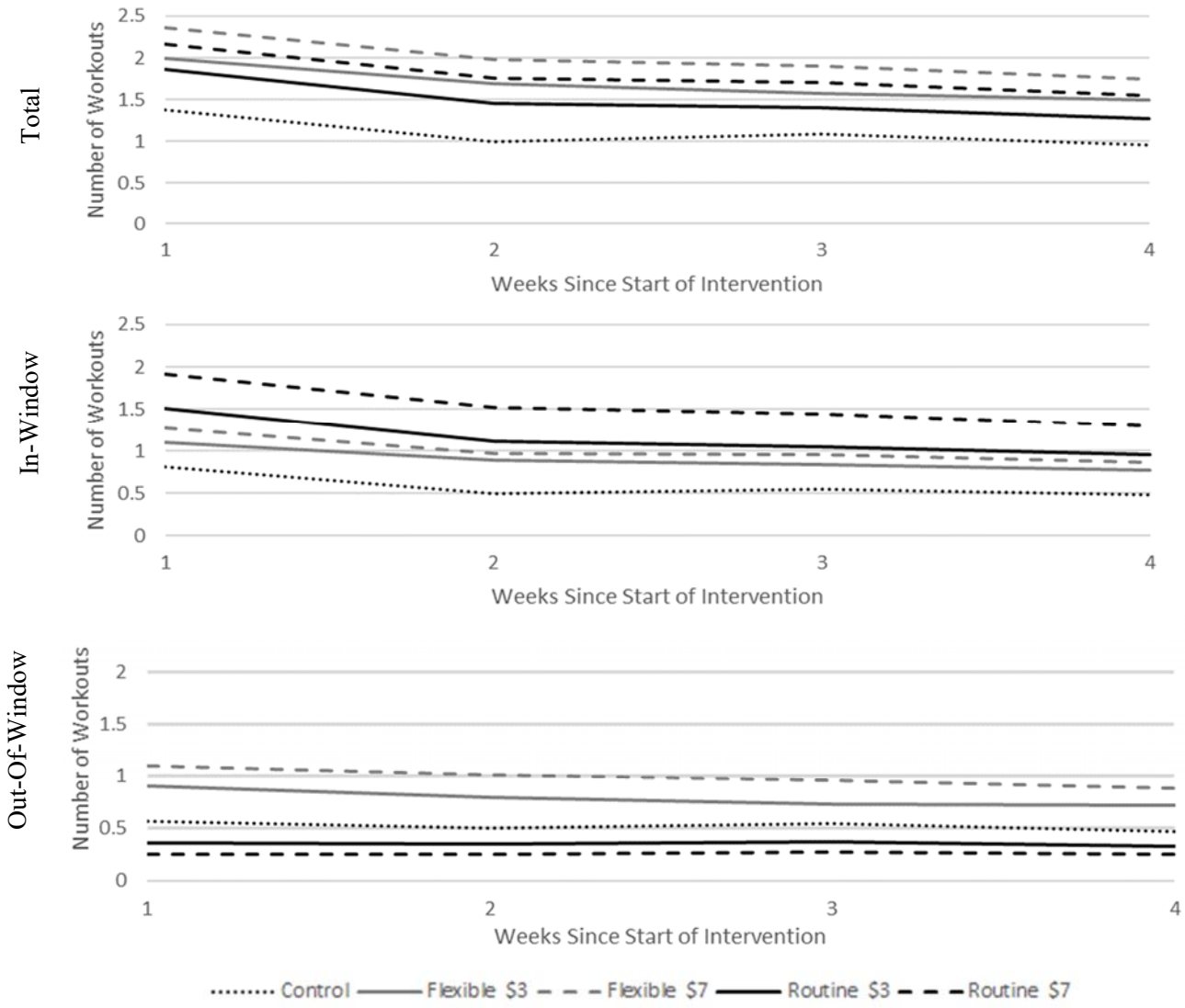


Figure 3. Panel B. Average Likelihood of Any Weekly Workouts for Four Weeks During our Intervention by Condition
This figure shows the average likelihood of working in a given week overall, in-window and out-of-window during our 4-week intervention period by experimental condition.

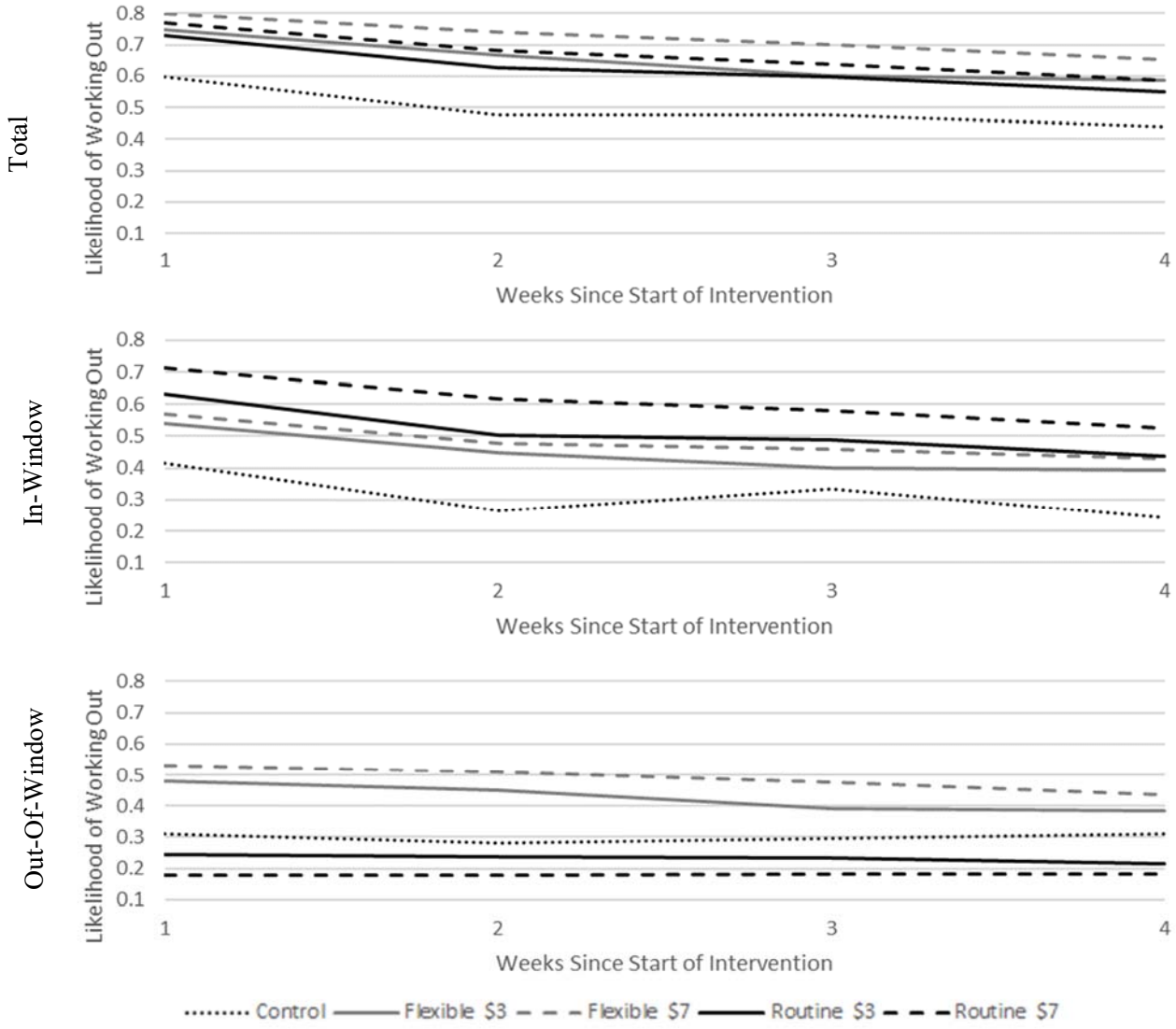


Figure 4. Panel A. Average Number of Workouts for Four Weeks Post-Intervention by Condition

This figure shows the average number workouts in a given week overall, in-window and out-of-window during the 4-weeks post-intervention by experimental condition.

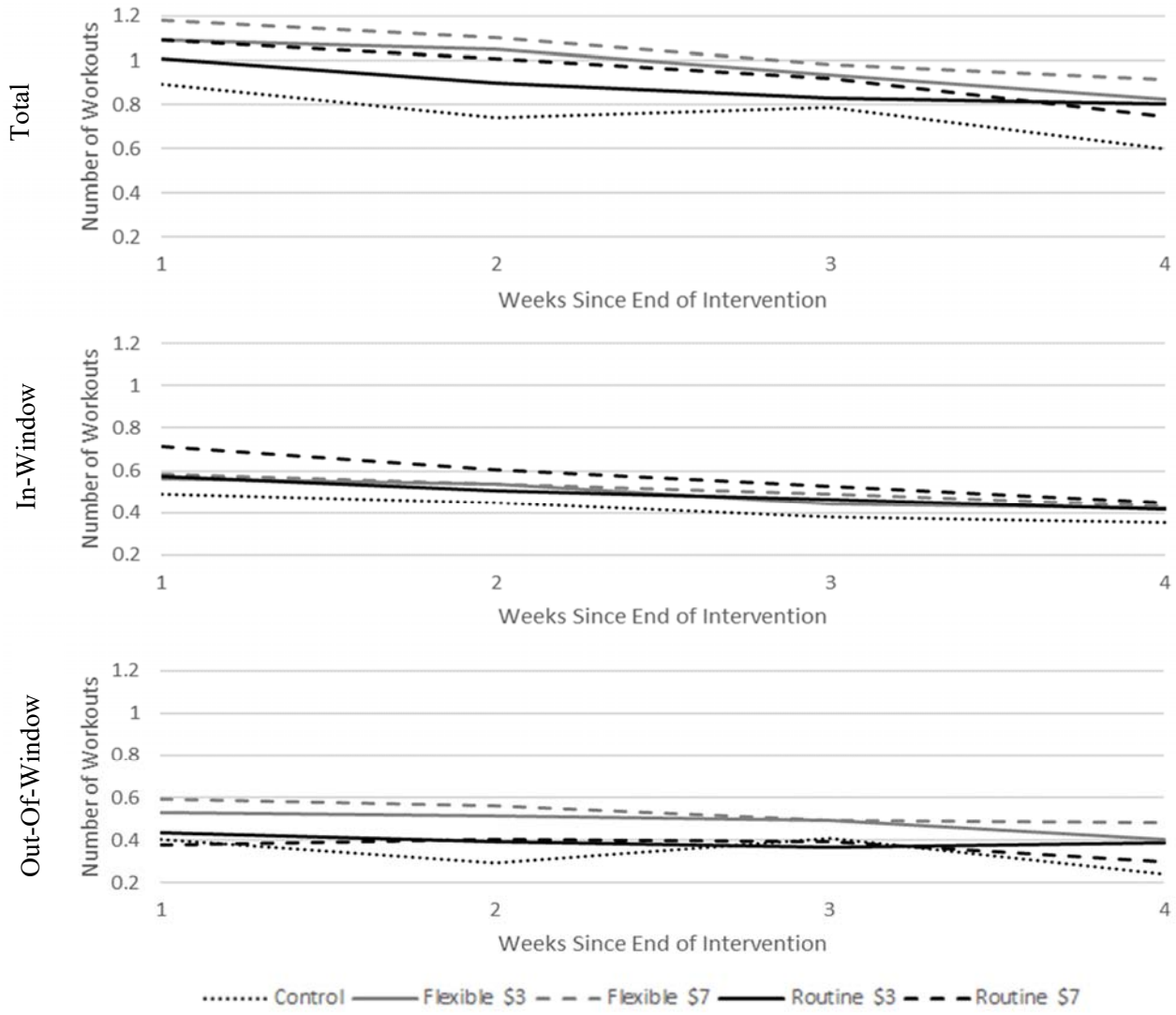


Figure 4. Panel B. Average Likelihood of Any Weekly Workouts for Four Weeks Post-Intervention by Condition

This figure shows the average likelihood of working in a given week overall, in-window and out-of-window during the 4-weeks post-intervention by experimental condition.

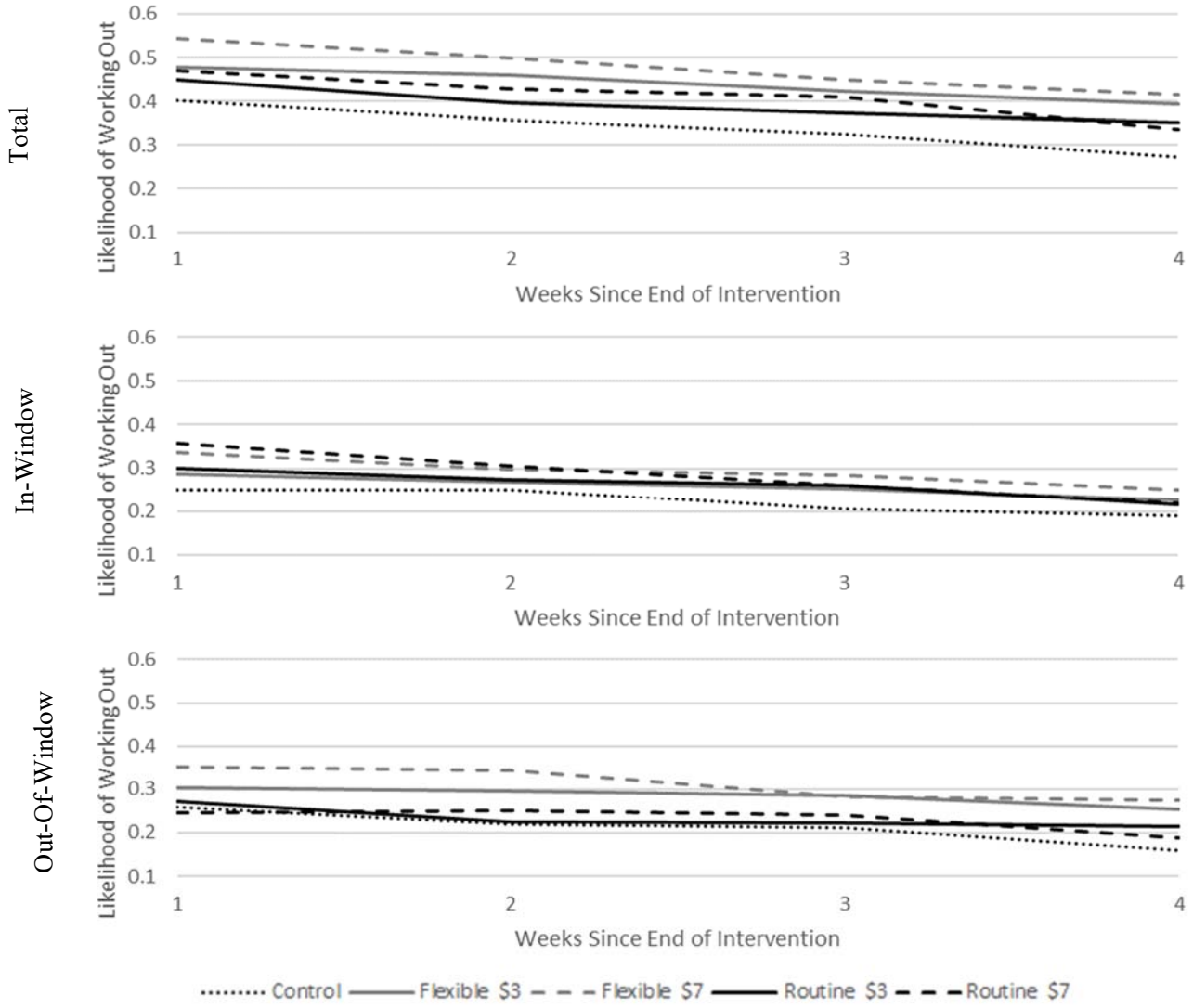


Figure 5. Comparing Exercise in the *Routine $7* Payment and *Flexible $3* Payment Conditions Pre- and Post-Intervention

This graph zooms in on two experimental conditions that induced similar numbers of *total* workouts during our intervention period, but the workouts these experimental conditions induced during our intervention were differently distributed: the *routine $7* payment condition produced many more in-window workouts and fewer out-of-window workouts than the *flexible $3* payment condition. As this figure depicts, while *routine $7* generated more workouts during our intervention period, this trend flipped post-intervention. For added visual clarity, we depict a 10-week post-intervention period in this graph.

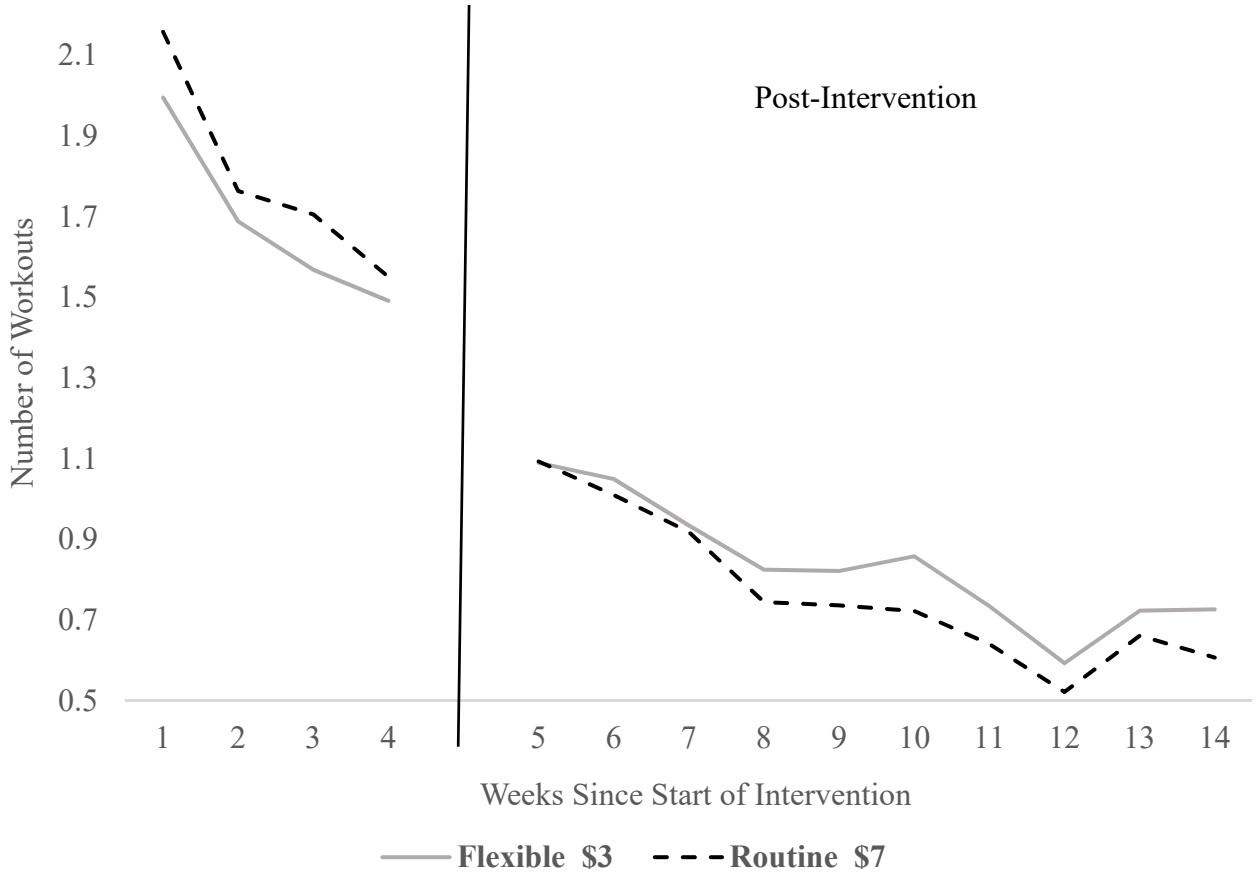
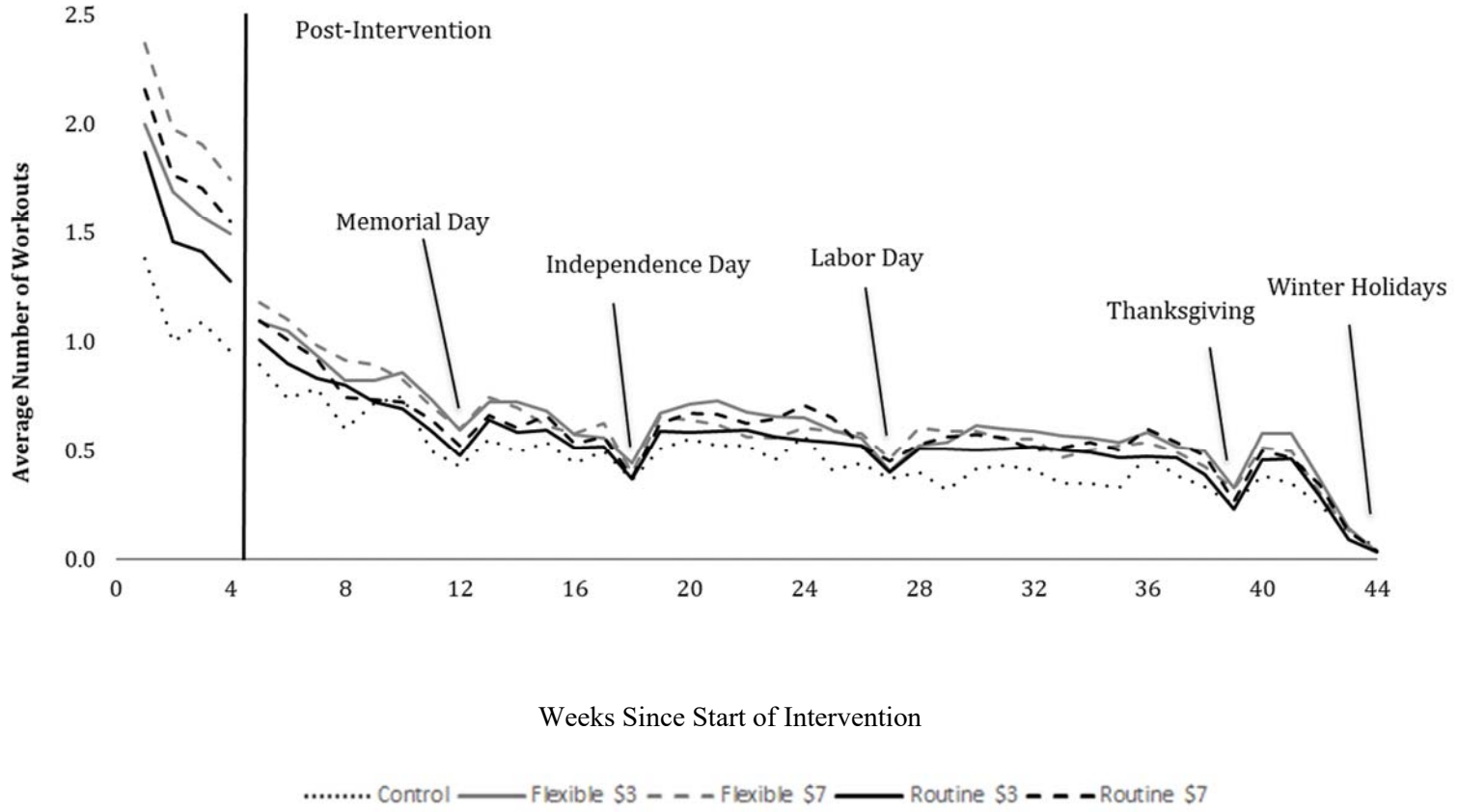


Figure 6. 40-Week Post Intervention Period

This figure shows the total average number of workouts in a given week during the 4-week intervention period, and 40-weeks post-intervention by experimental conditions. We labeled holidays, as they explain the reason we see dips in the average number of workouts.



APPENDIX

Our complete, 194-page appendix is available online at this website: http://www.opim.wharton.upenn.edu/~kmilkman/Gym/Appendix.pdf