NBER WORKING PAPER SERIES

INCENTIVES, COMMITMENTS AND HABIT FORMATION IN EXERCISE: EVIDENCE FROM A FIELD EXPERIMENT WITH WORKERS AT A FORTUNE-500 COMPANY

Heather Royer
Mark F. Stehr
Justin R. Sydnor

Working Paper 18580
http://www.nber.org/papers/w18580

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2012

We are thankful for funding from the National Science Foundation, the Upjohn Institute, and the Case Western Reserve University ACES fund. Royer also thanks the RAND corporation for support through the NIA. We are appreciative for the outstanding research assistant work of Stephen Cabrera, Andrew Chang, Vishal Chauhan, Tina Chen, Jon Evans, Natalie Greene, Brian Jameson, Victor Marta, Rachel Smith, and Bert Wagner. We appreciate the comments and suggestions of Nava Ashraf, John Beshears, Eric Bettinger, Tanguy Brachet, John Cawley, David Clingingsmith, Stefano DellaVigna, Uri Gneezy, Dean Karlan, Nicola Lacetera, and Jason Lindo along with those of various seminar and conference participants. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Incentives, Commitments and Habit Formation in Exercise: Evidence from a Field Experiment with Workers at a Fortune-500 Company

Heather Royer, Mark F. Stehr, and Justin R. Sydnor

NBER Working Paper No. 18580
November 2012, Revised August 2013
JEL No. D03,D9,I1

ABSTRACT

Financial incentives have been shown to have strong positive shortrun effects for problematic health behaviors, but the effects often disappear once incentive programs end. This paper analyzes the results of a largescale workplace field experiment to examine whether selffunded commitment contracts improve the longrun effects of incentive programs. Consistent with existing findings, workers responded strongly to an incentive targeting use of the company gym, but longrun effects were modest, at best. However, workers in the treatment arm that combined the incentive program with a commitment contract option showed longlasting behavioral changes, persisting even 1 year after the incentive ended.

Heather Royer
Department of Economics
University of California, Santa Barbara
2127 North Hall
Santa Barbara, CA 93106
and NBER
royer@econ.ucsb.edu

Mark F. Stehr
Drexel University
LeBow College of Business
Matheson Hall 504E
3141 Chestnut Street
Philadelphia, PA 19104-2875
stehr@drexel.edu

Justin R. Sydnor
University of Wisconsin - Madison
975 University Avenue
Madison, WI 53706
jsydnor@bus.wisc.edu

An online appendix is available at:
http://www.nber.org/data-appendix/w18580

Many people state a desire to change their behavior, yet struggle to do so. Common examples include desires to exercise more, save more money, or eat healthier food. These challenges have helped to motivate a rich literature in economics exploring models of time-inconsistent behavior.[1] This literature shows that present-bias can lead to consistent patterns of behavior that individuals perceive as suboptimal from their long-run perspective (O'Donoghue and Rabin, 1999, 2001).

The stakes involved with these time-inconsistency problems are particularly high in the case of health behaviors since they can have important long run consequences for quality of life and longevity. These issues are especially important in the US since American lifestyles are characterized by poor diet and a lack of physical activity.[2] The consequences also likely extend beyond the "internalities" that an individual's short-run self imposes on her long run self and generate important externalities as well. These unhealthy behaviors likely impact others through higher group-rated health insurance costs and increased spending on programs such as Medicare and Medicaid (Finkelstein et al., 2009).

In the face of these problems, there is increasing interest from individuals, firms, insurance companies, policy makers and health professionals in using financial incentives to motivate changes in health behaviors (Volpp, Pauly, Loewenstein and Bangsber, 2009; Baicker, Cutler and Song, 2010). The issue of incentives in health is currently pertinent for policymakers given the expanded scope that the Patient Protection and Affordable Care Act gives employers to use financial rewards and penalties to target health behaviors and outcomes.

A small literature has emerged to explore the effect of incentive programs on changing health behaviors (Volpp et al., 2008; Volpp et al. 2009; Charness and Gneezy 2009; Acland and Levy, 2011; Babcock and Hartman, 2011; Babcock et al, 2011; Cawley and Price, 2011; John et al., 2011). While this literature has its limitations, including sometimes small samples, specific populations, and issues with sample attrition, overall it points to strong responses to financial

[1] See Strotz (1955-56), Phelps and Pollak (1968), and Laibson (1997) for foundational work on time inconsistency in economic models of discounting. See also Frederick, Lowenstein and O'Donoghue (2002) for a review.

[2] As of 2009, only 14% ate the recommended amounts of fruits and vegetables (Seehttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC2654704/). According the Centers for Disease Control, in 2010, only 20.4% of adults met the CDC's muscle-strengthening and aerobic exercise recommendations. See http://www.cdc.gov/nchs/fastats/exercise.htm

incentives. However, these studies also often find disappointing long-run results where individuals fall back to old patterns of behavior once incentive programs end (Gneezy, Meier, Rey-Biel 2011). Understanding whether incentive programs can be designed to have more long-run effects is an important open question.

In this paper we present the results of a large-scale randomized field experiment testing the effectiveness of financial incentives for inducing lasting changes in exercise frequency in a working population. The experiment involved 1,000 employees at a Fortune 500 company and was conducted over two years. The treatment group was offered a one-month financial incentive to attend their company's onsite exercise facility ($10 per visit for up to 3 visits each week). The literature on time inconsistency would predict that this type of program could generate lasting change by helping those who were procrastinating to overcome the start-up costs associated with beginning to use the gym.[3] And in fact, broadly similar incentive programs have been shown to generate increases in exercise frequency for undergraduate populations lasting a few months (Charness and Gneezy 2009; Acland and Levy, 2011).

Yet incentive programs in other contexts have not shown lasting effects (e.g., John et al., 2011), and many people fail to use gym memberships in ways that suggest on-going problems of time inconsistency (DellaVigna and Malmendier, 2006). In light of those concerns, the primary innovation of this paper is a novel twist aimed at improving the lasting effect of incentives. After completion of the incentive period, half of the incentive group was randomly selected and offered the opportunity to create a self-funded commitment contract. This commitment contract allowed participants to put money at stake for a pledge that they would continue to use the gym over the 2 months following the original incentive period. If the employee kept to the commitment, she kept her money, but if not, the money was donated to charity. The incentive may kick-start behavior change, while the commitment option can potentially address ongoing challenges to maintaining that behavior.

We observe a strong response to the incentive program, with gym attendance doubling during the incentive period. A supplementary analysis of possible substitution – i.e., whether

[3] Projection bias, a further aspect of time inconsistency, where an individual exaggerates the extent to which his future tastes will resemble his current tastes, may compound the problem of overcoming initially high costs of a new exercise routine and could be helped by the incentive program (Loewenstein, O'Donoghue and Rabin, 2003).

these effects are true increases in overall gym attendance or a change in the location of exercise (e.g., from a non-corporate gym to the corporation gym) – suggests that while some substitution does exist, at least 70% of this treatment effect is new exercise.

After the incentive program ended, we find some lasting behavior change for those who were not members of the gym prior to the experiment, but overall the effects of the incentive program alone faded quickly. In the first month after the incentive, only 25% of the increase in exercise frequency persisted and by the second month, most of this increase was gone.

In contrast, the program pairing the incentives with the commitment-contract option successfully generated lasting changes in exercise frequency. Over the initial two months after the incentive ended (when the commitment contracts were in place), the group offered the combination of incentives and commitment retained half of their incentive-induced increase in exercise, attending the gym 50% more frequently than the control during this period. The effects for this group are very long-lasting effects - detectable even a year after the end of the incentive program.

These results show that commitment contracts can be a promising new way of improving the lasting response of an incentive program for exercise. Our work adds to a small but growing literature that has shown that various commitment technologies can be successful in promoting savings (Ashraf et al., 2006; Benartzi and Thaler 2004; Beshears et al., 2011; Giné et al., 2012), exercise (Milkman, Minson and Volpp, 2012), smoking cessation (Jeffrey et al., 1990; Gine, Karlan and Zinman, 2010) and weight loss (Jeffery et al., 1990; Volpp et al., 2008; John et al., 2011).[4] Our results also indicate that commitment contracts might be used in conjunction with periodic incentives as a cost-effective alternative to continuous incentives. At a more general level, the results broadly suggest that directly addressing the challenges of maintaining behavioral change may be an important direction for future work.

This study also provides some new insights about the demand for commitment contracts. Overall 12% of the employees offered the commitment option decided to take it, and among those who had attended the gym at least once during the incentive period, the take up rate was 22%. Our exploration of commitment demand revealed several interesting

[4] Goldhaber-Feibert, Blumenkranz, and Garber (2010) explore whether the commitment contracts people design for exercise can be influenced by anchoring and nudges but do not observe the outcomes of those contracts.

patterns related to demographics. Women, middle-age to older-age employees, and those who are overweight or obese are all more likely to create commitment contracts than their counterparts. Perhaps most interestingly we find that the demand for commitment is similar even among those who were exercising regularly prior to the study and have no apparent need for commitment. Prior studies of commitment contracts of the type used here have targeted populations with clear behavioral issues (e.g., smokers, obese), and as far as we know this is the first study that has explored the demand for commitment among those for whom there is no clear indicator of a potential behavioral problem. The demand for a likely non-binding commitment contract by those with high rates of exercise prior to the study suggests to us that the value of commitment devices may extend beyond their ability to change behavior affected by time inconsistency. For example, it may be that financial commitment contracts can substitute for other forms of self-control, which may have important welfare consequences if self-control results from the exertion of a limited supply of willpower (Baumeister et al., 1998, 2000; Ozdenoren, Salant and Silverman, 2012). Finally, we explore how proxies for the level of over-optimism about future exercise behavior relate to the demand for commitment. Although the literature on time inconsistency has tended to focus on the demand for commitment by "sophisticates" who are perfectly aware of (and not over-optimistic about) their level of present-bias, we find suggestive evidence that some degree of over-optimism may increase the demand for commitment.

2. Experimental Design and Data

2.1 Subject recruitment

The experiment took place at the headquarters of a Fortune 500 company located in the Midwest. At this location, there are approximately 1,900 employees holding a variety of jobs. The headquarters has a fitness center located on site that has the usual amenities of a modern gym. In order to use the gym, employees must become members of the wellness center and pay a membership fee of $12.96 every 2 weeks that is automatically withdrawn from their

paychecks.[5,6] Upon entry to the gym, employees log in at a computer terminal and these computerized log-ins serve as our primary data.

We began the experiment in February 2009 and enrolled our last participants in March 2011. We ran the experiment in 15 waves, with modest-size cohorts, to ensure that the gym staff could accommodate new gym member signups and that our results were not specific to a particular time of the year. Appendix Figure 1 describes the timeline of the experiment. We detail the number of participants along each step of the experiment in Appendix Figure 2.

To recruit subjects for each cohort, we first randomly drew a sample of employees from the company's full list of employees at the headquarters site, excluding high-level executives, human resource members, and gym staff privy to the details of the research. Although they knew that the field experiment involved incentives, the gym staff did not know who was participating in experiment. Then we sent the employees an invitation via e-mail to participate in two online wellness surveys (initial and follow-up) spaced 5 weeks apart. We described the experiment as a university study supported by the corporation. The employees were compensated with a $25 payment conditional on completion of both surveys. The initial survey collected a range of information on demographics, self-assessed fitness levels, exercise patterns, and subjective wellbeing. Response rates for this survey averaged 62% (see Appendix Figure 2).[7] We view this response rate as relatively high; for comparison, for the Card et al (2012) study of peer pay of UC employees, the survey response rate among employees was just above 20%. Subjects were informed that none of their individual responses to any surveys would be shared with anyone at the corporation. Since employees were aware they were participating in a study, this experiment is a "framed field experiment" (List, 2009).

Our pool of experimental subjects consists of the 1000 employees who responded to our initial survey. This even-number sample size was a random result of recruitment and not a targeted sample size. Upon completion of this survey, we randomized individuals into

[5] There are no start-up fees or contracts and employees can cancel their membership at any time with no penalty.
[6] The gym is open Monday through Friday from 6:00 a.m. to 7:00 p.m.
[7] Response rates do vary some across cohorts although in a regression of whether or not an individual responded on cohort fixed effects, we are unable to reject the hypothesis that the cohort fixed effects are jointly equal to each other. Moreover, the fraction of responders who are gym members is not changing systematically over time. If word spread rapidly through the company about the details of our experiment, we would expect that response rates and their fraction who are gym members would vary across cohorts.

treatment and control groups. The treatment group was eligible to receive financial incentives for gym attendance for a 4-week period whereas the control group was not; we elaborate on these treatments in more detail below. Because we anticipated that the response to incentives was likely to be heterogeneous, within each cohort we stratified the randomization into four groups: a cross of a) whether the subject was an existing member of the company gym and b) whether they responded in the initial survey that their current exercise was above or below their personal target frequency of exercise. After the completion of the incentive period, the incentive-eligible subjects were divided into two treatment groups (incentive and incentive+commit), detailed below. During the final week of the incentive program, all subjects who responded to the initial survey (including the control group) were asked to complete our follow up survey. This survey largely asked the same questions as the initial survey (omitting demographics). The response rate to this survey was 91.4% (see Appendix Figure 2).

Since the subject pool was not a random sample of all employees but rather consisted of individuals who responded to the initial survey, caution is warranted when extrapolating our results to the broader population of employees.[8] In light of the response emails we received, we suspect that a significant fraction of non-response was driven by those who traveling away from work during our recruitment. Of course, a company-sponsored program would not face these types of problems associated with communicating via email. Since we feel the observable characteristics we have for non-responders are unlikely to adequately characterize selection, we are reluctant to use these variables to predict what treatment effects would have been for the overall population. Instead we would argue that those interested in extrapolating population effects from our experiment might want to use the conservative approach of assuming that survey non-responders would not respond to financial incentives. At the end of our experiment, we contacted non-responders to our initial survey and assigned them to different treatment arms without having to fill out the initial survey. The response to the direct financial incentives for this subpopulation was small. Since our survey response rates are rather high, assuming no effect for non-responders would not change our conclusions qualitatively if extended to the full population.

[8]Our data on employees are limited (essentially departmental unit, position, and gym membership status). Gym members responded to the initial survey at a somewhat higher rate than gym non-members – 74% versus 57%.

2.2 First-level treatment: Financial incentives

Incentive-eligible participants could earn $10 for each visit (up to 3 visits per week and only 1 visit per day) to the corporate wellness center over a specified 4-week period. The treatment group also received a free gym membership during the incentive period (a value of $25.92). Additionally, since joining the gym involves a 1-hour new membership assessment, we offered $20 to new members to join. Since all treatment groups included both per-use incentives and the membership reimbursements/bonus, while the control group received neither, the incentive program is a package of incentives.[9] To ensure that the incentives were salient to participants, we informed treatment subjects via both email and via a physical letter sent via company mail. Based on evidence from follow-up surveys, lack of information about the incentive program was not an impediment to participation.

We measure gym use via the login records described above. As is common at most gyms (including in previous research on exercise incentives), the gym only uses a log-in process and does not require individuals to log out when leaving. As such, it is not possible to know how long the employee exercised or the nature of that exercise. In theory, there is some scope for employees to cheat on the program by logging in and not exercising, but our research assistants, who we asked to discretely monitor the gym, reported no such behavior. In addition, the gym staff -- who were aware of the program but did not know who was offered incentives -- reported no increases in suspicious logins and did not observe increases in employees showing up at the gym without exercising. Additionally, while such behavior could in theory be a concern during the incentive period, our primary interest is behavior after the incentive program ends, when the incentive for this cheating was much smaller.

Much of the interest in health-incentive programs to date has focused on incentivizing weight loss. For this study, we decided to incentivize gym-attendance rather than weight loss for several reasons. Most importantly, our interest in this study is in understanding how incentives interact with *behaviors* in situations where time inconsistency may matter. Exercising less often than one desires is a standard example of a behavior that may result from time inconsistency. Weight-loss, in contrast, is a desired *outcome* that could be achieved through a

[9] In pilot experiments at the company prior to this experiment, there was essentially zero response to a treatment offering only a free membership.

range of behaviors (some of which, e.g., use of diuretics, are unhealthy). Another reason for our focus on gym attendance is that while reducing rates of obesity is an important goal of health-promotion, there are clear and direct benefits to physical activity itself, including improved cardiac health, mental health, productivity, etc. Furthermore, the benefits of exercise are important to the broad population, both the obese and non-obese, which fits well with company-wide health promotion efforts. Fryer (2010) has made the point – in the context of educational incentives – that in general incentivizing positive behaviors may be more effective than incentivizing outcomes in situations where the production function mapping inputs to outcomes is not clear, which is likely the case for the health production function. Finally, it is possible in an experimental setting to observe gym attendance in a non-obtrusive way, whereas studies focusing on weight-loss generally require repeated weigh-ins and often suffer from high levels of attrition (e.g., Cawley and Price 2011).

2.2. Second-level treatment: Self-funded commitment contract

At the end of the 4-week incentive period, members of the treatment group were randomized into a second-level treatment, in which roughly half of the incentive eligible subjects were offered the chance to create a commitment contract. We refer to these two groups as the incentive-only and incentive+commit groups.[10] Up until the commitment contract offer, we treated these groups the same. Throughout, incentive+commit denotes the group *offered* the commitment option, and is an "intention-to-treat" grouping. The commitment contract for this study was a pledge not to go more than 14 calendar days in a row without attending the company gym over an 8-week period. Participants who decided to create a commitment contract could put as much money as they wanted towards the commitment. Commitments were self-funded, with participants placing their own money at stake with no external financial rewards. Subjects who successfully completed their commitment were

[10] In order to ensure balance between the incentive-only and the incentive+commit groups, we re-randomized during this step until a p-value on the test of the equality of the in-treatment effects between the two incentive groups exceeded 0.10. For the first few cohorts, we made these random sub-treatment assignments prior to observing exercise behavior from the incentive period. Given the relatively small sample size of cohorts, we observed some imbalance in gym visits between the incentive-only and incentive+commitment groups during the treatment period. For that reason we decided to change the protocols and conduct the randomization after the incentive period for later cohorts.

returned their money. In the case of a failed commitment, the committed money was forfeited to the United Way. To ensure an active response showing either interest or no interest to the commitment offer, the offer of a commitment contract was made when subjects were asked for their mailing address for their gym incentives and survey payment. Individuals who committed no more than they were owed for survey completion and gym-attendance simply risked receiving a reduced check from the experiment. Individuals could also commit more than they earned in the incentive program by writing a check made out to the United Way that was held until the end of the commitment period and returned if they successfully completed the commitment. Importantly, all payments for the gym-attendance incentive, including those for the incentive-only group, were mailed after this 8-week commitment period, so a subject who decided to create a commitment contract would not see a delay in receiving his or her incentive payment.

In order to keep the program simple so that it could be described briefly in an email and to reduce administrative burdens, we used a fixed commitment and did not allow for subjects to set the level of attendance for their commitment contract. The low attendance target was set such that it would be a reasonable minimum goal for anyone trying to exercise consistently and would be attractive to those most on the exercising margin. From an administrative perspective, this level of commitment also would not be too ambitious for employees with work-related travel or vacation, which usually extends less than a week at a time. Naturally, having a fixed contract with a modest goal likely made the contract less desirable to some participants, and it's possible that another contract would have performed better. Although we think that understanding optimal commitment contract design is an interesting and important area, we leave it for future research.

Subjects in the incentive-only group were sent a nearly identical email that encouraged them to commit themselves to not missing more than 14 days in a row at the gym over the following 8 weeks. This email did not, however, mention putting money at stake for that goal. Thus, the difference during the commitment period between the incentive-only and incentive+commit groups measures the effect of the offer of commitment rather than the combined effect of the encouragement and offer of commitment.

2.3. Data

Table 1 provides the means for key variables from our initial survey.[11] The table is split in two panels by gym membership status prior to treatment. Columns (1) and (4) show means for the control group with standard deviations of continuous variables for the control group in parentheses. To explore whether randomization provided balance in these characteristics across the different groups, we also display estimated mean differences between the control and incentive-only group (columns (2) & (5)) and between the control and incentive+commit group (columns (3) & (6)).[12] The last two columns in each panel are the p-values from two tests: first, the equivalence of the means across the 3 randomized groups and second, the equivalence of the means across the incentive-only and the incentive+commit groups. Overall, the groups are fairly well balanced across the different treatments; none of the pre-treatment differences examined in Table 1 are statistically different from zero at the 5% level.

Our subject pool is on average 40 years old, roughly equally divided across genders, and is well-educated (more than 65% have a college degree or more). In comparison, overall in the United States in 2009, just under 30% of adults aged 25 and older had at least a college degree. Possible time constraints are measured by marital status, presence of children at home, and commute times. Although marital status and presence of children at home are comparable to overall US patterns, commute times are significantly longer.[13] Company employees are on average somewhat less unhappy than in the US as a whole (14.3% report being unhappy in the 2010 General Social Survey).[14] Based on self reports of height and weight, 69% of our subjects are either obese or overweight, statistics that resemble those at the national level.[15] Both existing members of the gym and non-members report on average being around 20 lbs. heavier than their personal target weight.

[11] Note the sample sizes are not balanced across the three groups – control, incentive, and incentive+commit. We wanted the largest samples in the incentive and incentive+commit groups, which are approximately equal in size because their differences would be most difficult to detect.

[12] These estimated mean differences come from simple regressions that include strata fixed effects (a combination of gym membership, exercise relative to target and cohort), which are included in all regressions throughout. Including strata fixed effects ensures that results are not biased by fluctuations across cohorts in the shares of employees randomly sorted into control and treatment groups.

[13] Baseline statistics for this and previous sentence based on authors' calculations using the 2010 Census.

[14] Source of statistic is http://sda.berkeley.edu/cgi-bin/hsda?harcsda+gss10.

[15] http://www.cdc.gov/obesity/data/adult.html.

We asked subjects in the initial survey to report their current exercise activities and their targets for how often they would like to exercise. The average difference between targeted and self-reported exercise is 1.5 days/week for gym members and 2 days/week for non-gym members, implying that individuals want to increase their exercise and that incentives for exercise may move them closer to their target level. Given diminishing health returns to exercise, those who are inactive are likely to reap the largest returns. In our subject pool, rates of inactivity are high even among the gym members, as evidenced by the large fractions of individuals reporting no exercise in a typical week. Thus, our subjects likely have much to gain from increased exercise.

3. Conceptual framework

The design of this study -- a temporary financial incentive potentially followed by the opportunity to create a self-funded commitment contract -- is motivated by insights from the economics literature on time inconsistency. Before presenting the analysis of our results, we briefly lay out the conceptual background behind our experiment.

Individuals seeking to engage in behavioral change often face high startup costs, which in the context of exercise include joining a gym and adopting an exercise routine and new schedule. These large startup costs can result in sub-optimal behavior, particularly among those with present-biased preferences or projection bias. Faced with high initial costs to change and long-run future benefits, an individual with present-biased preferences may procrastinate on making such changes (O'Donoghue and Rabin, 1999, 2001). Relatedly, an individual with projection bias may not appreciate that the costs of exercise (e.g., pain) are likely to fall over time and hence may underinvest in establishing an initially difficult exercise habit (Loewenstein, O'Donoghue and Rabin, 2003). Thus, a temporary incentive could in theory provide the kick start a person with time inconsistency needs to establish lasting behavior change.

However, an initial reduction in the cost of exercise may not be enough for sustaining change. Activities like exercise, with present costs and future benefits, can generate recurring struggles for individuals with present bias. For instance, DellaVigna and Malmendier (2006) find that most gym members did not use the gym very frequently. Most surprisingly, this pattern held true for long-established members whom one might have expected would have quit once

they established that they did not use the facilities very regularly. One promising avenue for overcoming the struggles of present-bias is through commitment technologies motivated from quasi-hyperbolic discounting models (Strotz, 1955-56; Laibson, 1997; O'Donoghue and Rabin, 1999, 2001). In these models, individuals discount future utility using both a standard exponential discount rate and a present-bias coefficient that generates time inconsistency. Commitment technologies can potentially help individuals overcome present bias that leads to consistently sub-optimal behavior by committing their future selves to certain actions.

Following O'Donoghue and Rabin (1999), theoretical and empirical discussions of commitment demand have heavily focused on the degree to which a present-biased individual is aware of her time-inconsistency. Those who are fully aware of their present bias and recognize that they will face similar present bias in the future are commonly termed "sophisticates." Sophisticates may demand commitment devices that influence their future behavior because their present-bias, left un-checked, will lead to sub-optimal behavior in the future. Those who are overoptimistic about their level of future present-bias, in the sense that they predict that they will be less present biased in the future, are referred to as "partial sophisticates." An individual who is very overoptimistic about her future level of self-control (e.g., a naïf) may not perceive a need for a commitment contract. However, those with non-extreme overoptimism may see commitments as desirable (e.g., some partial sophisticates) but will likely believe that weak commitments will change behavior more than they actually will (Bryan, Karlan and Nelson, 2010).

To summarize, the literature on time inconsistency makes a number of broad predictions relevant for our study. First, the temporary incentive alone should be most effective at changing behavior for individuals who might procrastinate on overcoming high start-up cost for initiating an exercise routine. In our context, that is likely to be employees who are not ex-ante gym members. Second, the quasi-hyperbolic framework predicts that demand for commitment comes from those with time inconsistency problems. Those who report exercising less than they want or who rarely use their gym membership, would be likely candidates for commitment contracts. In contrast, those already successful at attaining their exercise targets should not generally need commitment. Third, this framework also posits that

as compared with naïfs, full sophisticates will have a greater demand for commitment. However, neither the theoretical nor the empirical literature has extensively discussed how moderate levels of overoptimism about future behavior affect the demand for commitment.

4. Results

4.1 Graphical analysis

The three panels in Figure 1 graph the time series of the fraction of subjects with at least one visit each week to the company gym over time by treatment status. Each point in the figures is a four-week average of the fraction attending the gym at least once in the week. We combine the data for each cohort such that month 1 is the 4-week incentive period. Months 2 and 3 encompass the period of the commitment contract.[16] The graphs go out for a full year from the beginning of the treatment period.

Figure 1a shows the overall results. As we would expect from random assignment, all three groups (control, incentive-only, and incentive+commit) had similar pre-treatment patterns, with on average approximately 20% of employees attending the company gym at least once each week. Those attendance rates were approximately doubled for the two treatment groups during the incentive period, revealing that employees responded strongly to the incentive treatment on average. Since the incentive+commit group was not informed of the commitment contract option until after the incentive period ended, we should see similar patterns for the two incentive groups during the incentive period. Although there is some difference in the in-treatment patterns, the effects are broadly similar.

Our primary interest is in behavior in months two and after, once the incentive period had ended. Not surprisingly, both incentive groups reduce their frequency of exercise relative to their incentivized levels. However, the two groups have distinctly different post-treatment patterns. The group offered only incentives reduces visit frequency almost to their baseline, with only a small lasting increase in visit frequency relative to control. In contrast, the attendance frequency of the incentive+commit group, 12% of whom decided to create a

[16] There was a week between the week of the initial survey and the start of the incentives that new members could use to sign up. Visits for that week are excluded from this graph. Also, for some cohorts the commitment period ran to week 14 due to holidays, so month 4 in the graph sometimes includes one week (week 13) that was within the commitment period.

commitment contract, remains clearly elevated relative to both pre-treatment levels and the control group over time. The differences are especially strong during months 2 and 3, when the commitment contracts were in place. During those months, approximately 30% of the incentive+commit group attended the gym at least once per week, while the control remained at the 20% baseline and the incentive-only group fell from around 25% in month 2 to just over 20% in month 3. Over time the attendance rates of the incentive+commit group slowly fall, but remain clearly elevated even a year after the one-month incentive treatment.

Since the commitment-contracts were no longer in place after month 3, the lasting effect of the incentive+commitment treatment is particularly striking. It is difficult to know exactly what mechanisms underlie the long-run effect. One possibility is that exposure to the idea of commitment contracts causes some individuals to enact their own commitment strategies after our formal contract period ends. It could also be that true habit formation requires longer than the one-month incentive period and that the commitment option helps some individuals exercise long enough to form a lasting habit. If that is the case, the results here suggest that commitment contracts could be a useful tool for incentive programs targeting behavior change in situations when it is unclear how long it takes for habits to change.

Figures 1b and 1c present time series separately based on gym-membership status prior to the experiment, which was a variable on which we stratified the randomization. Figure 1b. shows the patterns for those who were existing members of the gym prior to our experiment. Prior to the treatment, substantial fractions of gym members had low use of the gym, with only approximately 60% of existing members using the gym at least once in an average week.[17] That fraction rose to 80% during the incentive period for both incentive groups. Following the end of the incentive program, the incentive-only group's visit frequency fell back to match that of the control by month 3, and shows no real lasting response to the incentive. In contrast, the incentive+commit group (23% commitment take-up) shows a lasting response to the incentive

[17] The fraction attending falls over time for the control group, which is not surprising in this subsample because a) restricting to existing members naturally results in some reversion to the mean and b) high percentages of subjects had incentive periods in the fall and spring, so that the post-treatment periods are composed somewhat heavily of summer months when attendance tends to be lower.

program. Their attendance rates are approximately 10 percentage points higher than the control during months 2 and 3 and fall slowly, reaching the control group levels by month 11.

Figure 1c. shows the patterns for those who were not members prior to the experiment. Overall the incentive program motivated 15-20% of employees who were not already users of the gym to attend. The incentive alone had a clear lasting effect for this group, with attendance rates a few percentage points above those of the control even a full year out. This suggests that for a modest number of employees the temporary incentive program generated a permanent shift in the use of the company gym. The long-run effects for the group offered incentives and commitment contracts are even higher relative to control. The incentive+commit group attendance exceeds that of the incentive-only group for the entire post-incentive period, but we also observe a random but small imbalance (not statistically-significant) in the response to the per-visit incentives between these two groups (despite identical treatment during the incentive period). Our regression results and robustness tests below suggest that there are statistically-significant long-run differences between the groups, even after accounting for the small differential in-treatment response to the incentives.

4.2 Regression framework

To quantify our results, we run regressions using data from the pre-incentive, incentive, and post-incentive periods. Our regression models are of the following form:

$$y_{itw} = \alpha_0 + \alpha_1(IO) + \alpha_2(IC) + \alpha_3(member) + \delta_0 in-treatment + \delta_1(IO) \times in - treatment + \delta_2(IC) \times in - treatment + \beta_0 early\ post - treatment + \beta_1(IO) \times early\ post - treatment + \beta_2(IC) \times early\ post - treatment + \gamma_0 late\ post - treatment + \gamma_1(IO) \times late\ post - treatment + \gamma_2(IC) \times late\ post - treatment + \mu_s + \pi_w + \varepsilon_{itw}$$

where y_{itw} is an outcome measure, such as an indicator for attendance, for subject *i* in incentive week *t,* and calendar (not experiment) week *w. IO* is a dummy variable for the incentive-only group, *IC* is a dummy variable for the incentive+commit group*, member* is an indicator variable denoting whether the individual was a member of the gym prior to the intervention, *in-treatment* is a dummy variable for the in-treatment period, *early post-treatment* is a dummy variable for the initial post-treatment period (weeks 5-13), and *late post-*

treatment is a dummy variable for the longer post-treatment period (weeks 14-52). Our pre-specified strata fixed effects upon which randomization was based, represented by μ_s are fixed effects for each exercise vs. target, ex-ante company gym membership status, and cohort combination, giving us 2 x 2 x 15 strata fixed effects. π_w are week fixed effects and we estimate separate week fixed effects for members and non-members. Since there are weekly observations on the same individuals, we adjust the standard errors for clustering at the individual level. When we consider the effects of members and non-members separately, rather than pooled as above, some of the terms in the regression above are of course collinear (e.g., membership status) and hence dropped from the regression.

The regression above combines the effects of the 4 time periods of interest – pre-intervention, intervention, early post-intervention (i.e., commitment period), and late post-intervention into 1 regression, allowing for concurrent comparisons of effects. α_0 measures the mean outcome for the control group in the pre-intervention period. Thus, α_1 and α_2 measure differences for the incentive-only and incentive+commit groups relative to the control group, respectively in the pre-intervention period and should be near 0 due to randomization. δ_0 measures the mean level of the outcome for the control group during the incentive period relative to its pre-incentive period mean. Our "in-treatment effects" are given by δ_1 and δ_2, which are difference-in-difference parameters measuring the extent to which differences in the mean outcome for the incentive-only and incentive+commit groups, respectively, between the intervention and the pre-intervention periods differ from the analogous difference for the control group. We refer to $\hat{\delta}_1$ and $\hat{\delta}_2$ as our estimates of the effect of the incentives for the incentive-only and incentive+commit groups. We expect their values to be very similar since these groups are treated differently only in the post-intervention period. β_1 and β_2, along with γ_1 and γ_2, are difference-in-difference parameters analogous to δ_1 and δ_2, except that they measure the extent to which differences in the mean outcome for the incentive-only and incentive+commit groups between the post-intervention and the pre-intervention periods differ from the analogous difference for the control group. Since in the post-treatment period, the incentive+commit group is offered the commitment contract whereas the incentive-only group is not, we interpret β_1 and γ_1 as the effects of the incentives on behavior in the early and

late post-treatment period, respectively and β_2 and γ_2 as the effects of the incentives and the commitment contract for the early and late post-treatment periods, respectively. Thus, β_2 - β_1 is the effect of the commitment contract offer during the commitment period and γ_2 - γ_1 is an analogous effect except during the post-commitment period.

4.3 Regression results

We present our main regression results in Table 2 following our regression framework above. The table presents results for the full sample (columns 1 and 2), for existing members of the gym prior to the experiment (columns 3 and 4) and for non-members (columns 5 and 6). For each sample split we present two columns of estimates based on two outcomes: any visit in a particular week and average number of weekly visits.[18] We use subject-week observations for these regressions and cluster the standard errors at the subject level. With this structure in columns 1, 3, and 5 the dependent variable is an indicator that takes on value of 1 if the subject attended the gym at least once in that week and zero otherwise. In columns 2, 4, and 6 the dependent variable is a measure of the number of visits the subject made to the gym in that week, ranging from 0 to 5.

The regression estimates confirm the patterns discussed above for Figure 1. We detect no significant differences across the three groups in pre-period visit patterns. In column 1 we see that the incentive-only and incentive+commit groups were 18 to 20 percentage points more likely to attend the gym in a given week during the incentive period than was the control group. That is a doubling relative to the 20% baseline for the control group. In Column 2, the incentives led to 0.56 to 0.68 increases in the number of visits per week during the incentive period, more than a doubling of the frequency of visits relative to the control baseline. At the bottom of the table we display p-values from tests of the equivalence of the incentive-only and the incentive+commit group coefficients in the pre-incentive, incentive, and early and late post-incentive periods. Since the groups were treated the same during the incentive period, we

[18] For ease of interpretation, we present OLS estimates of these regressions. We also estimated probit models to take into account the binary nature of the dependent variable, "any visit," and these models produced similar results. The weekly visits measure is also bounded between 0 and 5 and in principle it would be appropriate to use a model that takes into account the censored nature of that dependent variable. Again for ease of interpretation we present OLS results. Tobit estimates yield very similar conclusions to the OLS regressions.

expect the in-treatment results to be similar. We find that not only are the magnitudes of the estimates similar, we also cannot reject that the treatment effects are the same for these two groups during the incentive period.

The estimates from the early post-treatment section of the table show results for the period immediately after the incentive program (weeks 5-13) when the commitment contract option was available to the incentive+commit group. Consistent with the graphical results, we find that during the first two months after the incentive program, visit frequency is slightly elevated (0.03) for the group offered incentives only relative to control. When compared to the in-treatment effects, the results in column 1 show that those offered incentives alone retained 17% (0.03/0.18) of their increase in visit frequency relative to the control over these two months. In contrast, the effects were longer lasting for the incentive+commit group. The frequency of visits for the incentive+commit group was 9 percentage points higher than the control over this period (a 40% difference in attendance). The commitment period effect is 45% of the in-treatment effect. The effects for the incentive+commit group in the early post-treatment period are larger than and are statistically different from the analogous effects for the incentive-only group; p-values of equivalence tests are 0.002 and 0.03 for the any visit and number of visits outcomes, respectively, as shown at the bottom of the table.

We can compare these effect sizes to two recent studies with undergraduate populations, Charness and Gneezy (2009) and Acland and Levy (2011), that both offered one-month incentive programs to motivate students to use the campus gym with incentives of a similar magnitude to those here. The in-treatment incentive effects for Charness and Gneezy (2009) and Acland and Levy (2011) imply that the incentives increase attendance by 1.2 visits per week. Our estimates are more modest – 0.56 visits per week, suggesting that employees are less responsive to incentives than university students. The post-treatment effect for Charness and Gneezy (2009) is 0.59 whereas for Acland and Levy (2011), it is 0.26. Our estimate of post-treatment effects for employees offered only incentives is again substantially smaller at 0.11. Expressed as a ratio of the in-treatment effect, the observed post-treatment effects in our study for the incentive-only group are close to those in Acland and Levy and about half the size observed by Charness and Gneezy.

Unlike the studies with undergraduate populations, in our setting we are able to provide estimates of longer-run effects covering weeks 14-52. For this longer post-period, there are no statistically significant differences relative to the control for the group offered only incentives. The incentive+commit group show significant and statistically significant increases in gym use relative to control over the longer run. The estimates in columns 1 and 2 both show that the incentive+commit group had attendance 25% higher than the control over the long run.

One reasonable question is whether these effect sizes for the incentive+commit group are plausible given the design of our commitment option. At the bottom of Table 2 we show the commitment-contract take-up rate for those offered commitment, which was 12% overall. Not surprisingly, the commitment rates of members exceeded that of non-members. However, when excluding those who did not attend the gym during the incentive period, the commitment rates are similar; 24% for members and 21% for non-members. The IV estimates (i.e., the treatment effects on the treated) at the bottom of Table 2 are estimates of the effect of the commitment contract for the early post-treatment period using the random assignment of the commitment contract offer. These estimates control for in-treatment visits and use only incentive and incentive+commit observations. Given the structure of the commitment contract (attend the gym at least once in a two week period), a purely mechanical IV estimate on any visit for an individual who does not exercise at all at the company gym would be 0.5 assuming perfect compliance. The actual IV estimates are generally around 0.5, suggesting that the intention-to-treat effect sizes we observe here are broadly sensible.[19]

In columns 3 through 6, we present results separately for those who were and were not existing members of the gym prior to our study. All of the patterns discussed above for the graphical analysis bear out in the regressions as well.

For existing members we estimate modest but statistically insignificant increases in gym use during the initial post-treatment period for those offered incentives alone. For those offered commitments, however we see significant increases over that period relative to control. Consistent with the graphs, in the longer-run we estimate zero difference in visit patterns for

[19] Of course, a lack of success in fulfilling the contract, the fact that many of the people partaking in these contracts are already exercising at the company gym, and the encouragement the contract may provide individuals to exercise beyond its minimal requirements will cause these estimates to stray from 0.5

those who received incentives only. We find modest long-run effects for the members offered incentives and commitments, consistent with the estimates for the pooled sample, but these differences are not statistically significant with the reduced sample size of members only.

For non-members in the incentive-only group, we estimate statistically significant increases in visit attendance in both the early and later post-treatment periods. The effect sizes are very similar in both of these periods, suggesting that the incentive program had a permanent effect of transitioning 3 to 4 percentage points more of the non-members to gym users relative to the control. Compared to the in-treatment effects, around 25% of the new gym use effect due to the incentives for this group is retained in the long run. The response of non-members in the incentive+commit group to the incentives is somewhat stronger than those in the incentive-only group; the difference is statistically-significant for the early post-treatment period but not for the late post-treatment period. Non-members in the incentive+commitment group had a 9 percentage point increase in the fraction attending the gym relative to control in the initial post-treatment period, which declines to 6 percentage points over the longer run.

Of course, for this non-member population, one concern with the comparison of behavior for those offered incentives only versus those also offered commitments is the differential response (albeit not statistically significantly different from one another) to the incentive program between these two groups. To address such concerns, we have also run separate analyses where we control for in-treatment visits, either through matching on visit patterns during the incentive period or controlling for such patterns. We consistently find differences in the usage patterns between incentive-only and incentive+commit groups during the post-treatment period using these approaches. For example, controlling for whether or not the individual attended the gym for each week of the incentive program, our estimate of the early post-treatment effect of the commitment contract offer (relative to incentives alone) is a statistically-significant 0.04, very close to the 0.05 treatment difference observed in Table 2. Thus, the observed in-treatment differences between the incentive and incentive+commit group have little impact on our conclusions about the long-run effectiveness of the commitment contract for non-members.

4.4 Commitment-contract take-up

In this subsection we explore the correlates of the demand for a commitment contract. We are cautious in interpreting these regressions because they rely on non-experimental variation and were not pre-specified. Nevertheless, given that our commitment treatment extended the effect of the incentive program, and that little work has explored how theoretical predictions map into actual commitment demand, we think this analysis can be informative.

Overall among the 346 subjects in the incentive+commit group, 12.4% chose to make a commitment and on average these committers placed $58 at stake.[20] Among ex-ante gym members, the take-up rate of commitments was 23%. For those who were not members of the gym prior to the study, the overall take-up rate was 6%, but take-up was 21% for those making at least one visit during the incentive period. Although these take up rates are somewhat modest, they are in line with existing studies. For example, Gine, Karlan and Zinman (2010) saw 11% take-up of their smoking-cessation commitment device in the Philippines. Ashraf, Karlan and Yin (2006) had a 28% take-up of their commitment savings product. Sixty-three percent of those who created commitments in our study successfully maintained the commitment of not missing more than 14 days in a row at the gym.

In Table 3 we present regression results examining the correlates of commitment-contract demand. For this analysis, we restrict the sample to those subjects who were offered the commitment option (incentive+commit group) and stated in the follow up survey (conducted during the last week of the incentive program) that they had interest in using the company gym over the following weeks (67% of the sample or 231 subjects).[21] In this way we focus on those who had some possibility of committing, since (unsurprisingly) none of those without interest in using the gym decided to make a commitment. The overall take up rate of commitment in this group was 19%.

Panel A presents regression results predicting take-up for this sample. In each column we include controls for the frequency of gym visits during the treatment period, breaking by

[20] We observe too few commitment contracts to present any meaningful analysis of the size of the commitment individuals made, and focus instead simply on the take up decision.

[21] The survey with these measures was conducted before subjects learned about the commitment option.

quartiles of average weekly visits.[22] These controls account for any "house money" effects the incentive earnings might have on commitment demand and more generally control for incentive program effects. Unsurprisingly we estimate that those who did not attend during the incentive program are very unlikely to make a commitment contract. More interestingly, however, rates of commitment are highest among those who exercised regularly but not enough to earn the full incentive amount. Thus, we do not think that "house money" effects, which predict that commitment contract takeup would be monotonically increasing with average visits, can fully explain our commitment takeup. Conditionally on visits, ex-ante members create commitment contracts at a higher rate than non-members but this member and non-member difference is not statistically significant.

Column 2 introduces demographic controls from the pre-intervention survey: gender, age, children at home, college degree and overweight/obesity. Men are significantly less likely (17 percentage points) to make commitment contracts than women. We also find a large age effect. Employees in the bottom quartile of age (age < 33) are 15-20 percentage points less likely to make a commitment contract than older employees.[23] The presence of children at home and one's education level are not significant correlates of demand. Finally, we find that being overweight or obese, as ascertained from self-reported data, strongly predicts take-up.

Most of the results in columns 1 and 2 do not speak in any obvious way to the theoretical motivations for commitment demand from the time consistency literature referenced earlier. Exceptions to this statement are the findings that falling just short of the full incentive earnings (3 visits per week) and being overweight both predict higher levels of commitment demand. Both of those indicators could be rough proxies for the existence of present-bias, and as such these results are in line with theoretical predictions that problems with present-bias drive a demand for commitment.

On the other hand, other findings in this table suggest that even for those seemingly without time-inconsistency problems, there is a demand for commitment. In columns 3 and 4,

[22] Average ≥ 3 is the omitted quartile in the regressions.

[23] These results hold if we include dummies for different age quartiles. We present the simple comparison of "young" employees versus "non-young" employees to simplify the exposition. There are little differences in take up rate by age among the higher quartiles of age.

when the gym membership dummy is replaced with the frequency of gym attendance prior to the intervention, we see that there is no significant effect for lower levels of pre-study exercise and if anything commitment takeup is highest among high use members. This result is unexpected since they are already regular attendees and the commitment contract requirements are far below their usual attendance. On a similar note, in column 4, we observe that being below one's personal exercise target, another possible measure of time inconsistency, is again not statistically significant and is actually slightly negatively correlated with takeup. Together these results suggest an expanded view of the demand for commitment. That is, time inconsistency may not be the only reason that individuals demand commitment. We discuss this finding and its potential relevance for our understanding of self-control problems and commitment in more detail in the concluding section.

In addition to predicting that individuals who are time-inconsistent will demand commitment, the theoretical literature and much of the empirical literature that has followed suggests that sophisticates (i.e., those aware of their self-control issues) will demand commitment. In contrast, overoptimism should reduce the demand for commitment. We further investigate this prediction in Panel B. In this panel, we include the controls from colum 3 in Panel A and restrict the sample to those who exercise less than their reported personal target from the initial survey. This is a group who self-identifies as having a potential self-control problem. The take-up rate of commitment contracts for these 155 subjects was 0.17. We then explore whether 3 different measures that relate to the level of sophistication versus overoptimism about a self-control problem are predictive of demand for commitment. In column 1 we include a measure of overoptimism obtained from the initial survey. Our measure is based on subjects' reported likelihood that they would hit their target level of exercise over the coming month before they knew anything about the incentive program. We label those predicting a probability of hitting their target of 50% (the median response) or greater as "optimists." This belief likely reflects overoptimism, since no control subjects classified as "optimists" actually hit their exercise target over the month following the initial survey. Optimists, as measured in this way, are more likely to make a commitment, with an estimated coefficient of 0.10. In column 2 we incorporate a measure of relative optimism based on

answers in our second survey (conducted during the last week of the incentive period but before any subjects were offered commitment contracts) about expectations of exercise behavior in the month following the end of the incentive period. Here (over)optimists are defined as individuals who expect to use the company gym more after the incentives are removed and/or who believe they will obtain their ideal level of exercise in the post-incentive period. Although there is no statistically significant difference, our estimates again go in the direction that over-optimism increases the take-up of commitment. In column 3 we find that those reporting strong levels of self-control related to exercise are no less likely to create a contract than those who appear to be self-aware of a self-control problem. Taken as a whole, the findings in Panel B of Table 3 suggest that some degree of over-optimism rather than full sophistication likely increases demand for commitment.

Section 5. Robustness and Heterogeneity.

5.1 Substitution

Our results above show that there were meaningful and lasting effects (especially for the incentive+commit group) of incentives on attendance at the company gym. Correctly interpreting these estimates, however, requires understanding how much of the response is due to increases in exercise or changes in the location of exercise. For example, subjects might simply start exercising at the company gym as a substitute for their exercise elsewhere. Substitution is an important issue with most incentive programs, since most target a particular measurable behavior, but the degree to which substitution occurs for these types of programs is largely unknown.[24] To test for substitution, we use data from the follow-up survey, which includes questions about overall exercise and exercise at the company gym during the incentive period.

[24] At least one other study in this area has attempted to measure substitution, but the conclusions are unclear. Charness and Gneezy (2009) ask participants to fill out an exercise log. The log includes questions about overall exercise, exercise at a gym, and exercise outside of a gym. As they state, the self-reported data in their case do not seem to be reliable. For example, the effect on gym use for the main incentive group is 0.04 gym visits/week whereas that measured via administrative data is 1.22 university gym visits/week. The difference in these estimates could reflect considerable measurement error in the exercise logs or significant substitution (i.e., substitution of other gyms for the university gym).

Table 4 presents the relevant estimates for our substitution analysis. We do this for different groups defined by the stratification variables we used for the randomization (i.e., membership status and level of exercise relative to target level of exercise); in that sense, this analysis is pre-specified before the start of the experiment. We provide estimates for members and non-members separately at the top of the table (Panel A). The remaining panels delve further into possible heterogeneity, dividing the sample by whether an individual's pre-intervention overall exercise was below their target (Panel B) or at or above their target (Panel C). For the Panel B estimates, we expected that the estimated effects would represent mostly new exercise. We postulated that for individuals at or above their target level of exercise, the incentives would not lead to increases in overall exercise but substitution of the location of exercise, especially in the case of non-members since for many of them earning incentives would only require that they move their existing exercise to the company gym.

Since our measures of overall exercise, which include exercise outside of the gym, are self-reported, it is useful to assess how reliable the self-reports are. To do so, we compare treatment effect estimates using the self-reported exercise at the company gym versus the estimates using the computerized data. We combine the incentive-only and incentive+commit groups because we are examining the substitution effects during the incentive period when these groups had identical treatments.[25] The self-reported and computerized-record estimates of the incentive effects are very similar in most cases (except members who are above target), increasing our confidence in the overall exercise results discussed below. Comparison of control group means across the computerized and self-reported data shows that existing gym members tend to overstate how frequently they attend the gym. However, this measurement error appears to be consistent across the control and treatment groups, leading to little bias in estimated treatment effects using the self-reported data.

To assess the degree of substitution, we compare the treatment effects for overall exercise to those for company gym exercise using the self-reported data. If the two estimated treatment coefficients are the same (a ratio of the overall exercise effect to the survey gym exercise effect of 1.0), we would interpret it as indicative of no substitution. Focusing on the

[25] Sample sizes differ across regressions because of non-response to the follow-up survey; regressions estimated using the computerized gym data on just the sample of follow-up survey responders give similar estimates.

overall effects in Panel A, we see that this ratio is 82% for existing members and 63% for non-members (a weighted average effect of 70%). These overall figures mask some interesting and predictable heterogeneity that is evident in panels B and C – for those reporting low levels of exercise relative to their target, the incentives appear to have led to increases in overall rates of exercising. In contrast, for non-members at/above their target exercise level, there appears to be considerable substitution.[26] Taken literally, 74% of the effect is substitution for that group.

This substitution analysis is based on information from our follow-up survey. Although the response rates to that survey are high (91.4%), we did observe some statistically-significant differential attrition in survey response between the treatment and control groups among non-members as seen in Appendix Table 1. In this table, we display estimates from regressions of whether or not an individual responded to the survey as a function of treatment status. To address the possible non-response bias for non-members in such analyses, we estimate the degree to which non-response might affect our substitution estimates in Appendix A. The upshot from these analyses is that the degree of response bias is small – leading to a possible understatement of substitution by roughly 5-10%. Overall since roughly 70% of our subjects were below their target level of exercise, even if we assume minimal effects for those at or above target and some response bias, the program generated a real change in exercise behavior for the majority of our subjects, particularly among those who stood to reap the largest health benefits (i.e., those who exercise the least).

5.2 Potential for cross-contamination and spillovers

One of the challenges in conducting a randomized workplace intervention is that since workplaces are usually closed environments, subjects in the experiment will often interact with each other.[27] One can imagine that these interactions could affect our estimates via two mechanisms: "cross-talk" or the discussion of the experiment within the firm and "spillover" or

[26] Survey gym visits do not appear to be a good measure of actual gym visits for members at or above their target. While the reason for this mismatch is not clear, there is little evidence this group increased their overall exercise in response to the incentives.

[27] The standard treatment effects literature assumes the existence of the stable-unit-treatment-value (SUTVA) assumption (Cox 1958) and such cross-contamination effects would be a violation of this assumption.

the interdependence of exercise behavior among individuals. In this subsection we consider how these factors impact our conclusions and argue that their impacts are likely minimal.

Cross talk would pose a problem if it changed the type of individuals who decide to enroll in the study. Such cross-talk would likely become more pronounced over time as more individuals are recruited for the experiment. However, the response rate to our recruitment survey does not change systematically over time and neither does the fraction of responders who are gym members. Additionally, our treatment effects are stable across cohorts, again supporting the idea that the selection of individuals into the experiment is not changing much. Aside from selection bias, cross-talk might affect the behavior of the control group if increasing knowledge of incentives available to the treatment groups leads them to become discouraged. However, as we see in Table 2, the control group attendance does not change from the pre-incentive to the incentive period.

Two types of spillover effects are important: interactions among those who have received incentives and interactions between those who have received incentives and the control group. In related work among college students, Babcock and Hartman (2011) find evidence of only the former type of interaction, although only among best friends. In our context, the spillover effects are likely to be small as we see no evidence of the treatment effect varying with the fraction of the department incentivized nor does the control group show any change in behavior from the pre-intervention to the intervention period that is related to the fraction of the control group's department that is incentivized. Moreover, only 3% of the company employees were incentivized at any one point in time.

5.3 Heterogeneity of response to incentives

In Table 5 we explore the heterogeneity in the response to our treatments by examining treatment effects for a number of different sample cuts. Except for the below target vs. at/above target split, these cuts were not pre-specified, so the results should be interpreted with caution. We motivate our heterogeneity analysis from the results on commitment take-up in Table 3, which suggested interesting patterns of commitment demand related to gender, weight, and age. Surprisingly, we also found evidence that those with high levels of pre-study exercise and those reporting being at/above their personal target were no less interested in the

commitment option than their low exercise or below target counterparts. In this section we briefly examine how our treatment effects compare for these various groups.

The results in Table 5 are broadly in line with what one would expect given the analysis in Section 4. Specifically, across all groups, the long-run effects are strongest for the incentive+commitment group. Moreover, these reduced-form long-run effects for the incentive+commitment group are generally consistent with commitment contract takeup. For example, across sex there are much larger differences between the incentive-only and the incentive+commit groups in the post-treatment period for women than for men. That result aligns with the fact that women make commitment contracts at much higher rates than the men. Interestingly, as an aside, while men and women responded fairly similarly to the incentive program during the treatment period, only men and not women show a lasting response to the incentive-only treatment. We also find statistically significant post-treatment effects of the incentive+commit group for the older employees and not for the younger employees, which is again consistent with the difference in their patterns of take-up of commitment contracts. Taken together, these results provide at least suggestive evidence in support of our findings in Section 4 that stronger effects of the incentive+commit treatment in the post period were driven by the availability of the commitment contract.

The heterogeneity table also provides an interesting look at the variations in treatment effects based on self-control problems related to exercise. Column 3 confirms our discussion from Section 4 that those who were exercising regularly prior to the study – as indicated by either exercise relative to target or pre-study exercise frequency -- created commitments at rates similar to their lower-exercise counterparts.[28] The post-treatment effects reveal, however, that the benefits of making commitments available are concentrated among those who were not consistent exercisers prior to the study. Although regular exercisers decided to make commitments, it does not appear that the incentive program or the availability of commitments altered their exercise habits in the post-incentive period. This lack of change is not surprising, given their high rates of exercise prior to the study, but it does highlight the

[28] We split subjects based on tertiles of pre-study exercise using gym attendance for existing members of the gym and self-reported exercise frequency in the initial survey for non-members.

intriguing nature of their decision to create commitment contracts despite no apparent prior need for, nor an apparent effect of the contracts on their behavior.

Section 6. Discussion and Conclusion

This study reports the results of a unique large-scale randomized incentive program targeting change in exercise behavior among a working population. We document that workers respond strongly to the incentive program while it is in place, but fall back close to pre-intervention rates of exercise quickly after the incentive is removed. Seen in that light, these results add to a large list of settings where health interventions have shown little ability to generate lasting changes in behavior. The primary innovation of this experiment, however, is to contrast this common temporary incentive approach with an alternative where participants are provided a self-funded commitment-contract option at the end of the incentive period. We find that the availability of the commitment option substantially improved the long-run effects of the incentive program. The incentive plus commitment program results in approximately a 50% increase relative to control in the fraction of employees exercising at the company gym during the two months following the end of the incentive. The effects are observable even a full year after the start of the incentive program, where we detect an increase of around 20-25% in the fraction using the gym relative to the control.

Our study provides a number of insights for organizations interested in using incentive programs to generate behavioral change. Sizeable fractions of working adults respond to a $10 per-visit incentive, which is a useful benchmark for employers wrestling with the decision to incorporate incentives into a broader wellness plan. However, we also find that relatively little of the money spent on incentives goes to new exercise. Taking into account pre-treatment exercise levels and our estimates of substitution effects, we conclude that approximately 35% of the cost of the incentive program was spent on new behavior, while 65% paid employees for exercise they would have done without the incentive. That in turn suggests that efforts to target incentive programs, when feasible, could be valuable.[29] More generally, the findings here suggest that programs incorporating temporary incentives with un-incentivized periods

[29] The targeting of incentives (e.g., payments for smokers to quit smoking) may be seen as inequitable and thus, while cost-effective, targeting may be rather infeasible.

that leverage habit formation through techniques such as commitment contracts will likely be more cost effective than consistent incentives.

Of course, this field experiment was designed to test behavioral responses to the incentive programs and not as an evaluation of a comprehensive workplace incentive program. We see this study as an important first step in understanding real-world incentive programs by employers, insurers, and other entities that aim to change health behaviors, but clearly more research is needed before we can speak to the optimal design of those programs. An employer-based program would likely involve a number of complementary efforts, company-wide communication, and probably a plan for at least periodic renewal of the incentives. Additionally, employers will be interested in assessing not only the behavioral response to their program but will also eventually hope to understand the extent to which these behavioral changes map into monetary impacts to the company through effects on absences, productivity, turnover, employee recruitment and health-care spending.

Although we cannot speak directly to these issues, we can provide some back-of-the-envelope calculations as a starting point for future studies. Baicker, Cutler and Song (2010) estimate that reductions in absences from work are a key channel for the benefits of workplace wellness programs and use a figure of $20 per hour (or ~ $160 per day) to value an absence. Based on that rate, the $57 per-employee cost of an incentive program like ours could be made up through reduced absences if it reduced the yearly per-employee absences by 0.36. That is, if roughly 1 in 3 employees experience one day less of absence due to the program, the program would be paid for through that channel. Alternatively, looking at health-care spending, the cost of our incentive+commit program was 1.1% of the $5,049 annual per-employee health-care cost reported by the Kaiser Family Foundation. Accounting for substitution, we estimate that our program increased the average frequency of exercise among employees by 26% over a one-year time frame. Thus, the program pays for itself in terms of reduced health care costs if a 26% increase in exercise frequency results in a 1.1% decrease in health care costs. Taken together, these back-of-the-envelope calculations suggest to us that a temporary incentive program coupled with a commitment contract to improve lasting effects would likely be cost effective for an employer.

This study also provides new insights for the literature on the use of commitment-contracts to address problems related to time inconsistency. In addition to our basic finding that commitment contracts can be used to improve the lasting effect of an incentive program, we also see a number of interesting patterns in the demand for commitment. We find that employees who were exercising consistently prior to the study make commitments at rates similar to those who appear more likely to have a time inconsistency problem. This is a unique finding of the paper that is possible only because ours is the first study designed to offer commitment contracts broadly to even those with no apparent need for commitment. One possibility consistent with this finding is that financial commitment contracts could serve as substitutes for the exertion of willpower in generating self-control. Research on the concept of ego depletion (Baumeister et al., 1998, 2000) and recent models of willpower depletion in economics (Ozdenoren, Salant and Silverman, 2012) suggest will-power is a depletable resource. It may be that having a financial commitment in place helps people to exert less personal effort to maintain self-control. In such situations commitment contracts might improve welfare even without measurably changing observed behavior.

We also find that among those who struggled to achieve their exercise target prior to the study, the demand for commitment was somewhat stronger for those who were more (over)optimistic about their future behavior. Much of the literature on commitment has focused on behavioral predictions under either complete awareness or complete unawareness of present bias (i.e., overoptimistic naifs). However, Bryan, Karlan, and Nelson (2010) highlight that the predictions of theory for the role of overoptimism are ambiguous. We believe our findings here suggest that understanding the role overoptimism plays in the demand for commitment and the effects overoptimism might have on the welfare effects of commitment programs should be an important goal for future research in this area.

References:

Acland, Dan and Matthew Levy. 2011. "Habit Formation and Naiveté in Gym Attendance: Evidence from a Field Experiment." Working Paper.

Ashraf, Nava, Dean Karlan and Wesley Yin. 2006. "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines." *Quarterly Journal of Economics,* 121(2): 673-697.

Babcock, Philip and John Hartman. 2010. "Networks and Workouts: Treatment Size and Status Specific Peer Effects in a Randomized Field Experiment" NBER Working Paper No. 16581.

Babcock, Philip, Kelly Bedard, Gary Charness, John Hartman, and Heather Royer. 2011. "Letting Down the Team? Evidence of Social Effects of Team Incentives" NBER Working Paper No. 16687.

Baicker, K., D. Cutler, and Z. Song (2010). "Workplace Wellness Programs Can Generate Savings." *Health Affairs,* 29(2): 304-311.

Baumeister, Roy, Ellen Bratslavsky, Mark Muraven, and Dianne Tice. 1998. "Ego Depletion: Is the Active Self a Limited Resource?" *Journal of Personality and Social Psychology*, 74(5): 1252-1265.

Baumeister, R. F., M. Muraven, M., and D.M. Tice 2000. "Ego Depletion: A Resource Model of Volition, Self-Regulation, and Controlled Processing" *Social Cognition* 18(2): 130-150.

Benartzi, Shlomo and Richard Thaler. 2004. "Save More Tomorrow: Using Behavioral Economics to Increase Employee Saving*." Journal of Political Economy*, 112(1): S164-S187.

Beshears, John, James J. Choi, David Laibson, Brigitte Madrian, and Jung Sakong. 2011. "Self Control and Liquidity: How to Design a Commitment Contract."

Bryan, Gharad, Dean Karlan, and Scott Nelson. 2010. "Commitment Devices." *Annual Review of Economics* 2: 671-98.

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

Cawley, John and Joshua Price. 2011. "Outcomes in a Program that Offers Financial Rewards for Weight Loss." In *Economic Aspects of Obesity*, Michael Grossman and Naci H. Mocan, editors, National Bureau of Economic Research.

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

Cox DR. 1958. *Planning of Experiments*. Wiley; New York.

DellaVigna, Stefano, and Ulrike Malmendier. 2006. "Paying Not to Go to the Gym." *American Economic Review*, 96: 694-719.

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

Finkelstein, Eric A., Justin G. Trogdon, Joel W. Cohen and William Dietz (2009). "Annual medical spending attributable to obesity: payer-and service-specific estimates." *Health Affairs*, 28(5):822-831.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue. 2002. "Time Discounting and Time Preference: A Critical Review." *Journal of Economic Literature*, 40(2): 351-401.

Fryer, Roland. 2010. “Financial Incentives and Student Achievement: Evidence from Randomized Trials.” National Bureau of Economic Research Working Paper No. 15898.

Giné, Xavier, Dean Karlan, and Jonathan Zinman. 2010. "Put Your Money Where Your Butt Is: A Commitment Contract for Smoking Cessation." *American Economic Journal: Applied Economics*, 2(4): 213–35.

Giné, Xavier, Jessica Goldberg, Dan Silverman, and Dean Yang. 2012. “Revising Commitments: Field Evidence on the Adjustment of Prior Choices.” Mimeo.

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

Goldhaber-Feibert, Jeremy, Erik Blumenkranz, and Alan M. Garber. 2010. “Committing to Exercise: Contract Design for Virtuous Habit Formation.” NBER Working Paper w16624.

Jeffery, Robert, Hellerstedt, Wendy L., and Schmid, Thomas L. 1990. “Correspondence programs for smoking cessation and weight control: A comparison of two strategies in the Minnesota Heart Health Program.” *Health Psychology*, 9(5): 585-598.

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

Laibson, D. 1997. “Golden Eggs and Hyperbolic Discounting.” *Quarterly Journal of Economics,* 112(2): 443–477.

List, John. 2009. “An Introduction to Field Experiments in Economics.” *Journal of Economic Behavior and Organization* 70(3): 439-442.

Loewenstein, George, Ted O’Donoghue, and Matthew Rabin. 2003. “Projection Bias in Predicting Future Utility.” *The Quarterly Journal of Economics* 118(4): 1209-1248.

Milkman, Katherine L, Julia A Minson, and Kevin G.M. Volpp. 2012. “Holding the Hunger Games Hostage at the Gym: An Evaluation of Temptation Bundling.” Working paper.

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

O’Donoghue, Ted and Matthew Rabin. 2001. “Choice and Procrastination,” *Quarterly Journal of Economics*, 116(1): 121-160.

Ozdenoren, Emre, Stephen Salant, and Daniel Silverman. 2012. "Willpower and the Optimal Control of Visceral Urges." *Journal of the European Economic Association,* 10(2): 342-368.

Phelps, E. and R. Pollak. 1968. “A second-best national saving and game-equilibrium growth.” *Review of Economic Studies,* 35: 185–199.

Strotz, R.H. 1955-56. “Myopia and Inconsistency in Dynamic Utility Maximization.” *Review of Economic Studies,* 23(3): 165–180.

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

Volpp, Kevin G., Mark V. Pauly, George Loewenstein, and David Bangsberg. 2009. “An Agenda for Research on Pay-For-Performance For Patients.” *Health Affairs,* 28(1): 206-214.

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

Figure 1a. Fraction with positive gym visits by treatment status (all subjects)

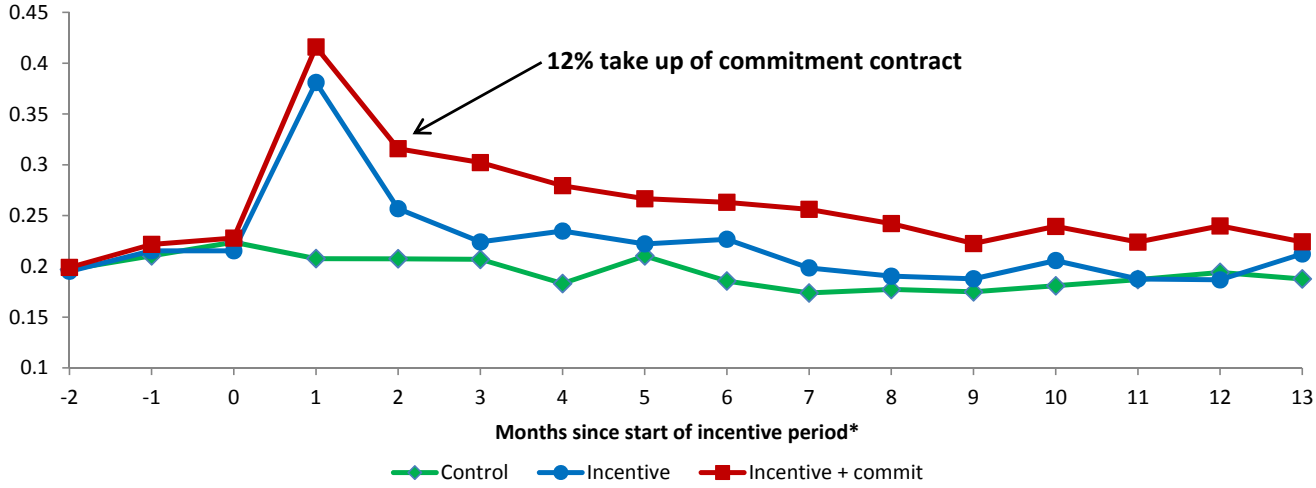


Figure 1b. Fraction with positive gym visits by treatment status (members only)

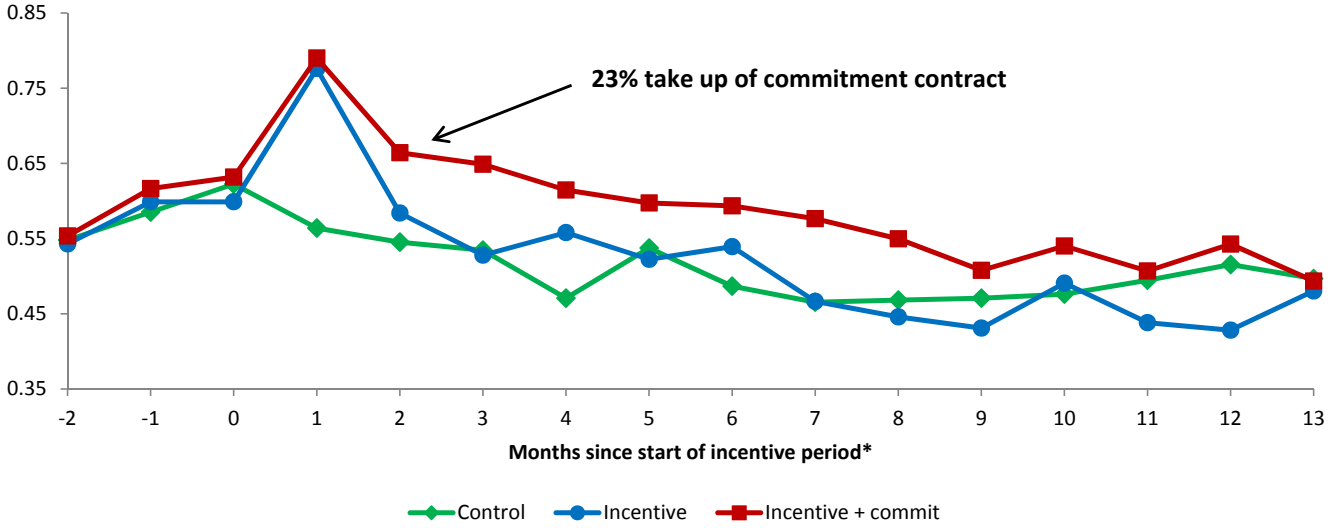
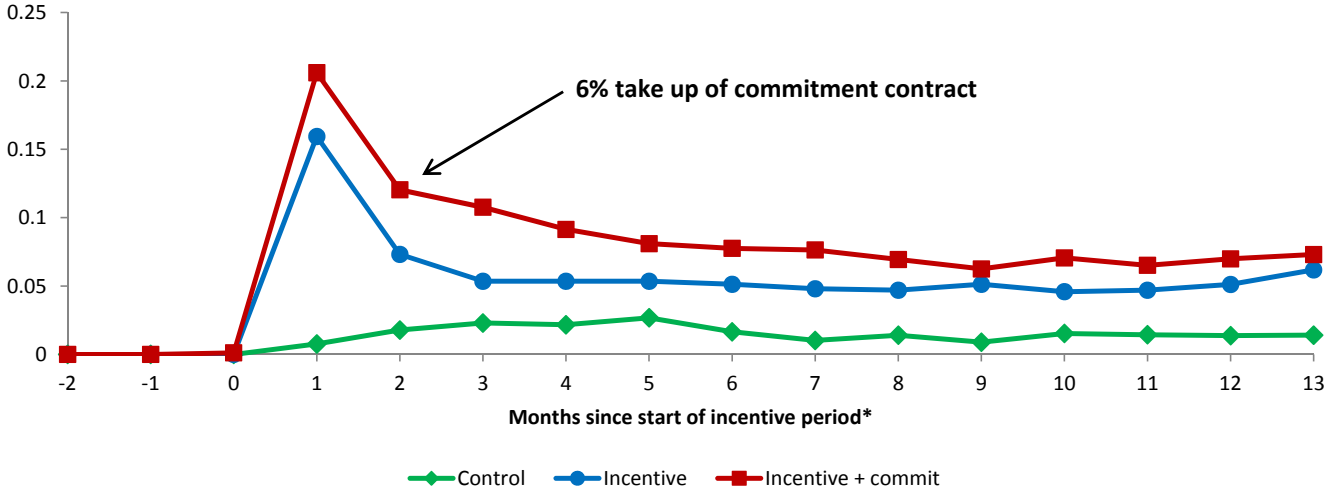


Figure 1c. Fraction with positive gym visits by treatment status (non-members only)



* Note: We definte months for this figure in 4-week blocks. Month 1 denotes the 4-week incentive period. Months 2 and 3 are the commitment-contract period.

Table 1. Pre-Treatment Descriptive Statistics

	Members					Non-Members				
	(1)	(2)	(3)			(4)	(5)	(6)		
	Control Mean	Incentive-Only Diff	Incentive+ Commit Diff	p-value (2)=(3)=0	p-value (2)=(3)	Control Mean	Incentive-Only Diff	Incentive+ Commit Diff	p-value (2)=(3)=0	p-value (2)=(3)
Basic Demographics										
Age	40.12	-1.17	-0.01	0.61	0.37	39.62	-0.75	0.16	0.65	0.36
	(10.63)					(10.82)				
Male	0.46	0.04	0.08	0.53	0.63	0.52	0.01	-0.01	0.90	0.71
College Degree or More	0.61	0.09	0.05	0.37	0.52	0.64	0.03	0.10	0.06	0.09
Living Situation										
Married	0.68	0.02	0.04	0.78	0.71	0.67	-0.03	0.01	0.65	0.31
Has at Least 1 Kid at Home	0.45	0.08	0.07	0.46	0.89	0.48	-0.05	0.04	0.19	0.06
One-Way Commute (Minutes)	37.82	-2.35	-0.21	0.48	0.28	38.03	0.76	-0.85	0.69	0.41
	(21.23)					(20.54)				
Subjective Wellbeing										
Unhappy with Life	0.07	0.00	0.02	0.89	0.70	0.11	0.01	0.01	0.91	0.86
Unhappy with Fitness	0.34	0.05	0.01	0.64	0.53	0.54	-0.10	-0.08	0.08	0.72
Unhappy with Weight	0.58	-0.01	-0.08	0.33	0.23	0.55	-0.04	-0.05	0.61	0.90
Health and Fitness										
Pounds over Target Weight	20.28	-1.63	1.42	0.69	0.37	22.72	-0.91	2.00	0.58	0.29
	(18.68)					(28.59)				
BMI	28.31	-0.60	-0.29	0.68	0.67	28.22	-0.49	0.40	0.36	0.13
	(5.52)					(6.51)				
Overweight	0.43	-0.04	0.03	0.50	0.29	0.42	-0.06	-0.04	0.44	0.62
Obese	0.30	0.00	-0.06	0.47	0.30	0.30	-0.02	0.03	0.43	0.18
Takes Blood Pressure Meds	0.12	0.03	0.00	0.78	0.57	0.13	0.00	-0.02	0.75	0.49
Exercise										
Average Days of Overall Exercise	3.36	-0.10	0.12	0.34	0.16	1.98	-0.13	-0.09	0.54	0.73
	(1.65)					(1.73)				
Target Days of Exercise	4.79	0.03	0.19	0.26	0.17	4.05	-0.18	0.00	0.30	0.20
	(1.08)					(1.33)				
0 Days of Overall Exercise	0.05	0.03	-0.03	0.16	0.07	0.24	0.01	0.02	0.86	0.76
Number of Observations	94	134	131			195	228	215		

Notes: Columns (1) and (4) are the control group means. Columns (2), (3), (5), and (6) are the mean differences between that group and the control group; these are estimated via regressions that include strata fixed effects. Standard deviations for continuous variables are presented with parentheses.

Table 2. OLS Regression results

Dependent variables: Any visit = 0/1 indicator whether individual attended gym in a given week
Weekly visits=number of visits an individual had in a given week

	Overall		Members		Non-Members	
	Any Visit	**Weekly Visits**	**Any Visit**	**Weekly Visits**	**Any Visit**	**Weekly Visits**
Control mean of dep var in pre-period	0.20	0.58	0.62	1.80	0.01	0.03
Incentive only	-0.01	-0.06	-0.02	-0.18	-	-
	(0.02)	(0.06)	(0.05)	(0.16)		
Incentive + Commit	0.01	0.00	0.01	-0.02	-	-
	(0.02)	(0.06)	(0.05)	(0.16)		
In-treatment period (weeks 1-4)	0.00	0.00	-0.02	-0.09	0.01**	0.05**
	(0.01)	(0.03)	(0.03)	(0.10)	(0.01)	(0.02)
(Incentive only) x (In-treatment)	0.18**	0.56***	0.23***	0.87***	0.15***	0.38***
	(0.02)	(0.06)	(0.04)	(0.13)	(0.02)	(0.06)
(Incentive + Commit) x (In-treatment)	0.20**	0.68***	0.21***	0.95***	0.20***	0.53***
	(0.02)	(0.07)	(0.04)	(0.13)	(0.03)	(0.08)
Early Post-treatment (weeks 5-13)	0.01	0.01	-0.03	-0.14	0.03***	0.09***
	(0.01)	(0.04)	(0.02)	(0.09)	(0.01)	(0.03)
(Incentive only) x (Post-treatment)	0.03*	0.11**	0.03	0.15	0.04**	0.09*
	(0.02)	(0.05)	(0.03)	(0.11)	(0.02)	(0.05)
(Incentive + Commit) x (Post-treatment)	0.09**	0.23***	0.10***	0.30***	0.09***	0.21***
	(0.02)	(0.05)	(0.03)	(0.10)	(0.02)	(0.06)
Late Post-treatment (weeks 14-52)	0.02	0.04	-0.01	-0.10	0.04***	0.12***
	(0.01)	(0.04)	(0.03)	(0.11)	(0.01)	(0.03)
(Incentive only) x (Post-treatment)	0.02	0.06	0.00	0.01	0.03**	0.09**
	(0.02)	(0.05)	(0.04)	(0.13)	(0.01)	(0.04)
(Incentive + Commit) x (Post-treatment)	0.05**	0.15***	0.04	0.14	0.06***	0.16***
	(0.02)	(0.05)	(0.03)	(0.12)	(0.02)	(0.05)
Subject-week observations	56,654	56,654	20,369	20,369	36,285	36,285
Number of subjects	1000	1000	359	359	641	641
P-values test of equal effects -- incentive-only vs. incentive + commit:						
Pre-treatment	0.37	0.21	0.47	0.26	0.26	0.26
In-treatment (weeks 1-4)	0.34	0.12	0.73	0.52	0.16	0.14
Early post-treatment (weeks 5-13)	0.002	0.03	0.03	0.15	0.02	0.09
Late post-treatment (weeks 14-52)	0.07	0.07	0.15	0.24	0.24	0.20
Commitment contract take-up:	0.12	0.12	0.23	0.23	0.06	0.06
IV estimate on weeks 5-13 attendance:	0.45***	0.59*	0.38***	0.58	0.65**	1.01
	(0.14)	(0.35)	(0.15)	(0.42)	(0.29)	(0.83)

Notes: Robust standard errors clustered by individual in parentheses. All regressions excluding the IV estimates include strata fixed effects (i.e., study cohort x exercise above or below target fixed effects) and separate week fixed effects. For the overall regressions, separate week fixed effects for members and non-members are included. The IV estimates are measured as the effect of a commitment contract on gym attendance using assignment to the incentive+commit group as an instrument. The included covariates in these IV estimates are fixed effects for the outcome for each week of the incentive period (e.g., number of visits in week 1 of incentive period for the visits dependent variable) and strata fixed effects. *** p<0.01, ** p<0.05, * p<0.1

Table 3: OLS Regressions Predicting Uptake of Commitment Contracts

Panel A. Those offered commitment who express desire to use the gym.
Dependent variable: indicator for whether subject made a commitment contract

	(1)	(2)	(3)	(4)
Mean takeup rate:	0.19	0.19	0.19	0.19
Avg weekly visits in-treatment: (avg = 0)	-0.12*	-0.15**	-0.12	-0.12
	(0.07)	(0.07)	(0.08)	(0.08)
Avg weekly visits in-treatment: (0 < avg < 2)	0.04	0.02	0.05	0.05
	(0.08)	(0.07)	(0.08)	(0.09)
Avg weekly visits in-treatment: (2 ≤ avg < 3)	0.15	0.17*	0.19*	0.19*
	(0.09)	(0.09)	(0.10)	(0.10)
Member prior to study	0.04	0.06		
	(0.06)	(0.06)		
Member prior to study: bottom tertile of use			0.03	0.03
			(0.08)	(0.08)
Member prior to study: middle tertile of use			0.05	0.05
			(0.08)	(0.08)
Member prior to study: top tertile of use			0.13	0.12
			(0.09)	(0.10)
Male		-0.20***	-0.20***	-0.20***
		(0.05)	(0.05)	(0.05)
Young (bottom quartile: age ≤ 32)		-0.13***	-0.13***	-0.13***
		(0.05)	(0.05)	(0.05)
Has children		-0.02	-0.02	-0.02
		(0.05)	(0.05)	(0.05)
College degree		0.03	0.03	0.03
		(0.05)	(0.05)	(0.05)
Overweight or obese		0.12**	0.13**	0.13**
		(0.06)	(0.06)	(0.06)
Below target for exercise in pre-survey				-0.02
				(0.07)
Number of subjects	231	224	224	224

Notes: Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1. Sample restricted to those in treatment group who were offered commitment option and further restricted to those who expressed a desire to exercise at the company gym in the post treatment period (231 of 347). The takeup rate among those with no desire was zero. Columns 3 and 4 control separately for members with different pre-treatment visit frequencies (base group remains non-members). See text for further details. Reduced sample size in columns 2-4 due to missing age for some observations.

Panel B. Sample from Panel A restricted to those below target for exercise pre-treatment
Dependent variable: indicator for whether subject made a commitment contract

	(1)	(2)	(3)	(4)
Mean takeup rate:	0.17	0.20	0.17	0.20
"Optimist-pre": believes high probability of hitting exercise target	0.10*			0.17**
	(0.06)			(0.08)
"Optimist-post": believes visit frequency will rise and/or hit ideal level in post-incentive period		0.06		0.04
		(0.08)		(0.08)
Self-reports "strong self-control" for exercise			0.03	0.00
			(0.06)	(0.09)
Controls from Column 3 of Panel A included?	Yes	Yes	Yes	Yes
Number of subjects	155	123	155	122

Note: Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1. Sample from Panel A is further restricted to those who reported exercising below target in the pre-treatment survey. All variables from Column 3 of Panel A are included here. "Sophisticated-pre" indicates a belief that there was less than a 50% chance (median reported chance) of hitting personal target for exercise over the month following the initial survey. The "low self-control" variable is based on median split to question in initial survey asking respondents to guage their own level of self-control for exercise on a likert scale. "Sophisticed-post" identifies those who stated on the initial follow-up survey administed before these subjects learned about the commitment option that they believed their use of the gym would fall after the treatment period and would be below their ideal level of use. Sample size in Columns 3 and 4 reduced to missing observations for exercise expectations.

Table 4a. Substitution Analysis for Members - In-Treatment Effects

Panel A: All subjects

Dependent Variable:	Weekly visits Gym data	Weekly visits Survey data	Overall exercise Survey data
Incentive-only or inc+commit	0.79***	0.49***	0.40**
	(0.15)	(0.17)	(0.18)
Observations	359	335	337
Mean control	1.59	2.26	3.25

Panel B: Subjects reporting exercise below their target in pre-survey

Dependent Variable:	Weekly visits Gym data	Weekly visits Survey data	Overall exercise Survey data
Incentive-only or inc+commit	0.88***	0.76***	0.76***
	(0.18)	(0.21)	(0.22)
Observations	209	190	192
Mean control	0.97	1.46	2.32

Panel C: Subjects reporting exercise at/above their target in pre-survey

Dependent Variable:	Weekly visits Gym data	Weekly visits Survey data	Overall exercise Survey data
Incentive-only or inc+commit	0.60**	0.03	-0.16
	(0.26)	(0.29)	(0.29)
Observations	150	145	145
Mean control	2.58	3.51	4.71

Table 4b. Substitution Analysis for Non-Members - In-Treatment Effects

Panel A: All subjects

Dependent Variable:	Weekly visits Gym data	Weekly visits Survey data	Overall exercise Survey data
Incentive-only or inc+commit	0.45***	0.60***	0.38***
	(0.05)	(0.07)	(0.13)
Observations	641	571	572
Mean control	0.03	0.07	2.09

Panel B: Subjects reporting exercise below their target in pre-survey

Dependent Variable:	Weekly visits Gym data	Weekly visits Survey data	Overall exercise Survey data
Incentive-only or inc+commit	0.47***	0.65***	0.45***
	(0.06)	(0.08)	(0.15)
Observations	499	446	447
Mean control	0.003	0.03	1.58

Panel C: Subjects reporting exercise at/above their target in pre-survey

Dependent Variable:	Weekly visits Gym data	Weekly visits Survey data	Overall exercise Survey data
Incentive-only or inc+commit	0.37***	0.43**	0.11
	(0.14)	(0.19)	(0.31)
Observations	142	125	125
Mean control	0.11	0.20	3.96

Notes: Dependent variable is average of weekly visits or exercise over the treatment period (i.e., one observation per subject). Regressions include strata fixed effects. Robust standard errors are reported. Cuts of the data above are based on our pre-specified randomization stratification by target level of exercise x membership status.

Table 5. Heterogeneity Cuts on Treatment Effects

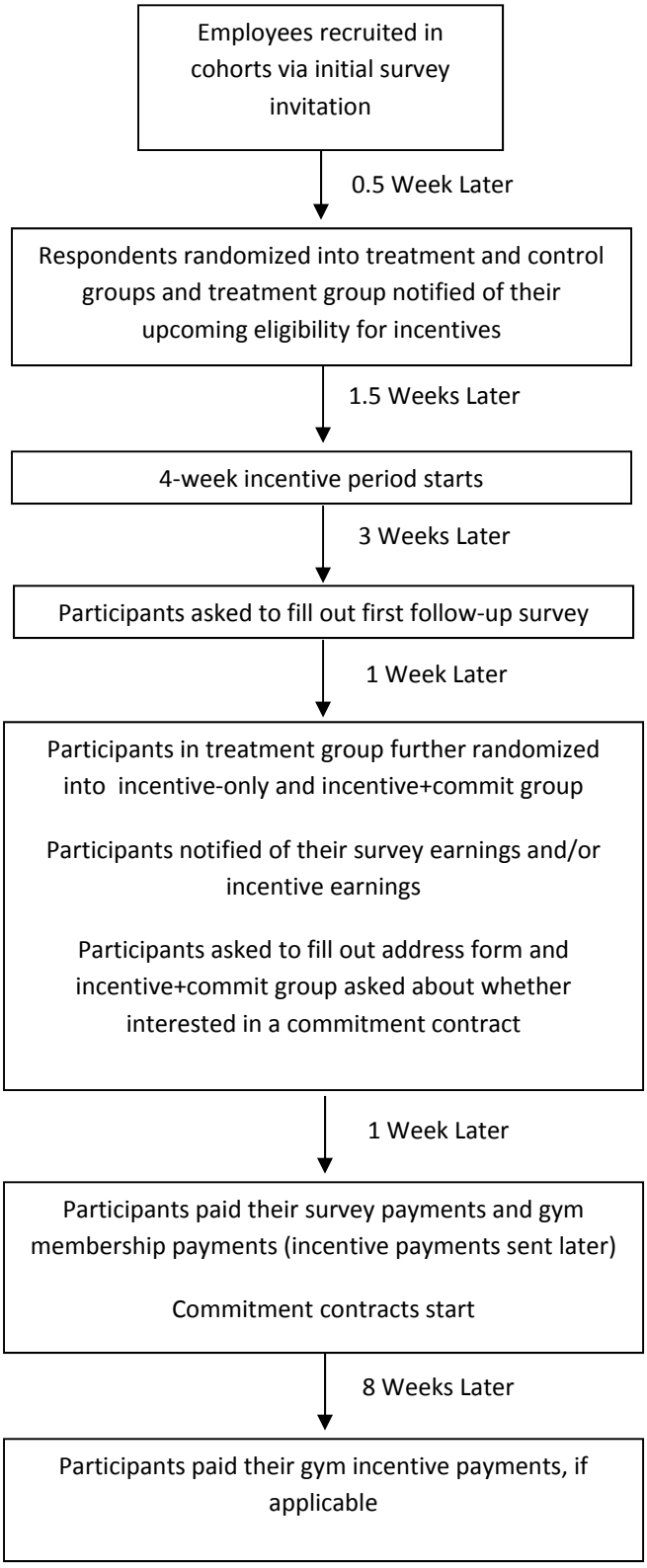
Dependent variable for columns (2), (4), and (5): Any visit = 0/1 indicator whether individual attended gym in a given week

		In-treatment effect		Early post-treatment effect (week 5-13)		
	Control mean (weeks 1-13)	Pooling incentivized groups	Commitment contract takeup	Incentive-only	Incentive+Commit	P-value on post-treat effect differences: (4) = (5)
Heterogeneity cut	(1)	(2)	(3)	(4)	(5)	(6)
Exercise						
Low pre-period exercise	0.03	0.17***	0.09	-0.00	0.07*	0.03
		(0.03)		(0.03)	(0.04)	
Middle pre-period exercise	0.11	0.23***	0.12	0.07**	0.13***	0.07
		(0.03)		(0.03)	(0.03)	
High pre-period exercise	0.41	0.11***	0.15	-0.02	0.03	0.17
		(0.03)		(0.04)	(0.04)	
Exercise relative to target						
Below target	0.11	0.21***	0.11	0.03	0.12***	0.001
		(0.02)		(0.02)	(0.02)	
At/above target	0.40	0.13***	0.15	0.01	0.05	0.28
		(0.03)		(0.04)	(0.04)	
Sex						
Female	0.22	0.17***	0.18	-0.02	0.09***	0.0004
		(0.03)		(0.03)	(0.03)	
Male	0.16	0.20***	0.07	0.07**	0.11***	0.12
		(0.03)		(0.03)	(0.03)	
Obesity						
Non-obese/overweight	0.23	0.18***	0.09	0.04	0.12***	0.04
		(0.03)		(0.03)	(0.04)	
Obese/overweight	0.17	0.18***	0.14	0.01	0.09***	0.0004
		(0.02)		(0.02)	(0.02)	
Age						
Below 33 years old	0.20	0.17***	0.05	-0.02	0.05	0.04
		(0.03)		(0.03)	(0.03)	
33 years old and older	0.18	0.19***	0.15	0.04	0.12***	0.001
		(0.02)		(0.02)	(0.02)	

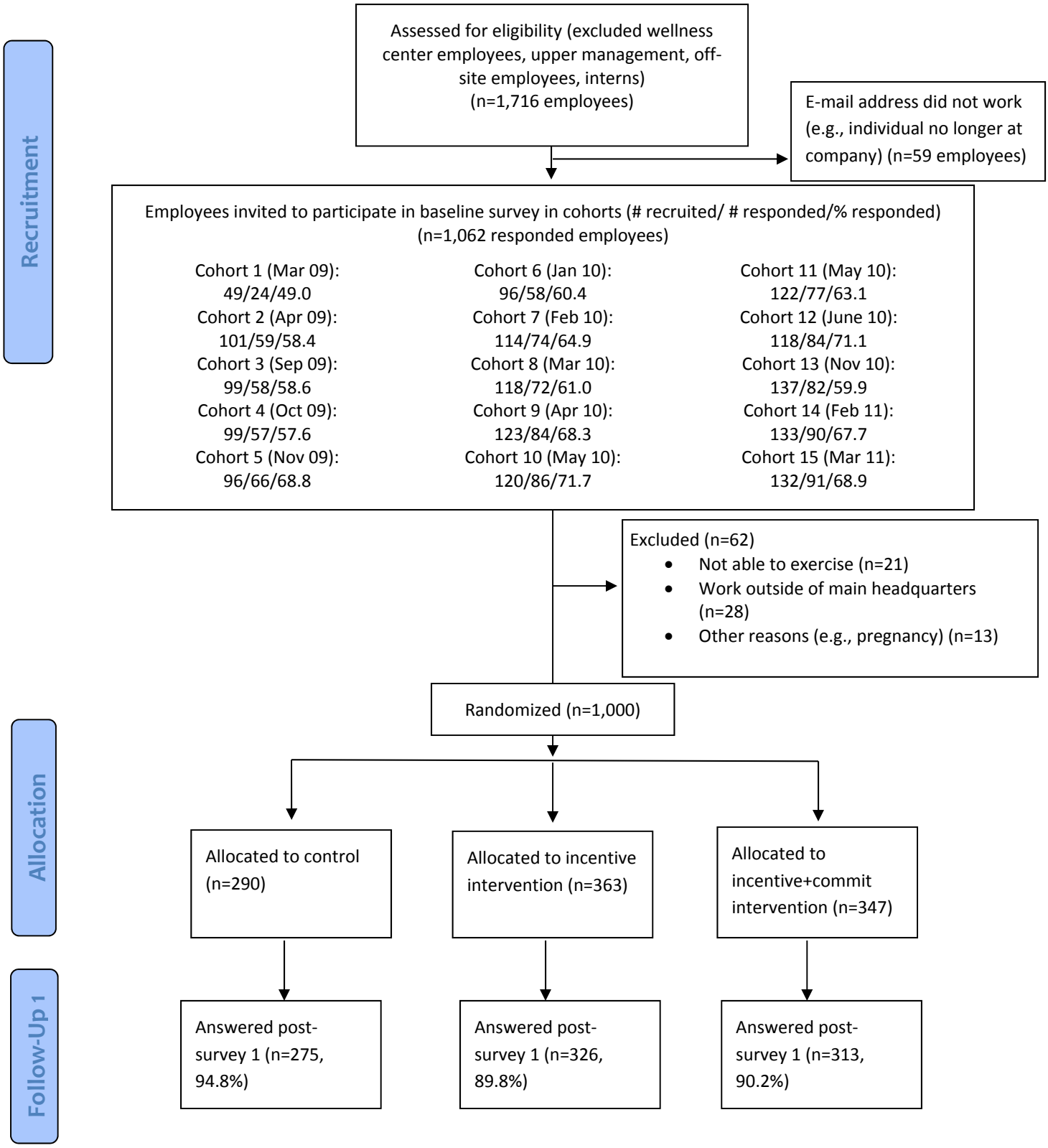
Notes: Each row of estimates is based on the sample indicated in the first column. Robust standard errors clustered by individual in parentheses. The reported coefficient in column (2) is the coefficient on a dummy variable indicating whether or not the individual was eligible for financial incentives during the incentive period (i.e., equal to 1 for the incentive or incentive+commit group and 0 otherwise). In columns (4) and (5), we report the coefficients from dummy variables indicating membership in the incentive group (column (4)) and membership in the incentive+commit group (column (5)). Strata fixed effects are included in all regressions. All regressions pool members and non-members. *** p<0.01, ** p<0.05, * p<0.1

FOR ONLINE PUBLICATION

Appendix Figure 1. Timeline



Appendix Figure 2. Flow Diagram



Appendix Table 1. Post-survey response rates as a function of treatment

Dependent variable: Indicator for whether subject responded to the post-survey

	Members			Non-Members		
	All	Below Target	Above Target	All	Below Target	Above Target
Mean for control group	0.93	0.91	0.94	0.96	0.96	0.95
Incentive-only or inc + commit	0.02	0.01	0.03	-0.09***	-0.09***	-0.09*
	(0.03)	(0.04)	(0.04)	(0.02)	(0.02)	(0.05)
Number of subjects	359	209	150	641	499	142

Robust standard errors in parentheses. All regression estimates control for strata. *** p<0.01, ** p<0.05, * p<0.1

Appendix A: The Substitution Effect after Accounting for Non-Response Bias

First, it is imperative to define the substitution effect. We define the substitution effect as the ratio of the effect of the incentives on overall days of weekly exercise to the effect of the incentives on days of weekly company gym exercise. A ratio of 1 indicates that there is no substitution whereas a ratio of 0 reflects complete substitution. Mathematically, we define the true substitution effect as the following:

$$\text{True substitution} = \frac{\overline{Y}^{overall}_{treated} - \overline{Y}^{overall}_{control}}{\overline{Y}^{gym}_{treated} - \overline{Y}^{gym}_{control}}$$

where $\overline{Y}^{overall}_{treated}$ is the average overall weekly days of exercise for the treated group (i.e., the incentive-only and incentive+commit group combined) during the incentive period, $\overline{Y}^{overall}_{control}$ is the average overall weekly days of exercise for the control group, $\overline{Y}^{gym}_{treated}$ is the average weekly days of company gym exercise for the treated group during the incentive period, and $\overline{Y}^{gym}_{control}$ is the average weekly days of company gym exercise for the control group. In contrast, the measured substitution effect from the post-survey data:

$$\text{Measured substitution} = \frac{\overline{Y}^{overall}_{treated,r} - \overline{Y}^{overall}_{control,r}}{\overline{Y}^{gym}_{treated,r} - \overline{Y}^{gym}_{control,r}}$$

where the means pertain to the group of survey responders (i.e., $\overline{Y}^{overall}_{treated,r}$ is the average weekly days of overall exercise for the treated group survey responders).

The measure of true substitution above can be expanded as follows:

$$\text{True Substitution} = \frac{p_{treated,r}\overline{Y}^{overall}_{treated,r} + (1-p_{treated,r})\overline{Y}^{overall}_{treated,nr} - (p_{control,r}\overline{Y}^{overall}_{control,r} + (1-p_{control,r})\overline{Y}^{overall}_{control,nr})}{p_{treated,r}\overline{Y}^{gym}_{treated,r} + (1-p_{treated,r})\overline{Y}^{gym}_{treated,nr} - (p_{control,r}\overline{Y}^{gym}_{control,r} + (1-p_{control,r})\overline{Y}^{gym}_{control,nr})}$$

where $p_{treated,r}$ is the probability of a treated individual responding to the survey and $p_{control,r}$ is the probability of a control individual responding to the survey. The nr subscripts on the averages denote non-responders (e.g., $\bar{Y}^{overall}_{control,nr}$ is the average weekly days of overall exercise for the control group non-responders). We can simplify this expression as $\bar{Y}^{gym}_{treated,nr} = 0$ and $\bar{Y}^{gym}_{control,nr} = 0$ because in our data, we observe that all non-responders did not attend the gym during the incentive period. Therefore, the true substitution effect simplifies to

$$\text{True Substitution} = \frac{p_{treated,r}\bar{Y}^{overall}_{treated,r} + (1-p_{treated,r})\bar{Y}^{overall}_{treated,nr} - (p_{control,r}\bar{Y}^{overall}_{control,r} + (1-p_{control,r})\bar{Y}^{overall}_{control,nr})}{p_{treated,r}\bar{Y}^{gym}_{treated,r} - p_{control,r}\bar{Y}^{gym}_{control,r}}$$

$$= \frac{\bar{Y}^{overall}_{treated,nr} - \bar{Y}^{overall}_{control,nr} + p_{treated,r}(\bar{Y}^{overall}_{treated,r} - \bar{Y}^{overall}_{treated,nr}) - p_{control,r}(\bar{Y}^{overall}_{control,r} - \bar{Y}^{overall}_{control,nr})}{p_{treated,r}\bar{Y}^{gym}_{treated,r} - p_{control,r}\bar{Y}^{gym}_{control,r}}$$

We have data on all objects in this formula except for $\bar{Y}^{overall}_{treated,nr}$ and $\bar{Y}^{overall}_{control,nr}$. One might consider a reasonable approximation for these objects to be their pre-incentive levels from the initial survey. Using these values, we can show that our estimates of true substitution and measured substitution are close to one another; they differ by at most 10%.