American Economic Journal: Applied Economics 2015, 7(3): 51–84
http://dx.doi.org/10.1257/app.20130327

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

By Heather Royer, Mark Stehr, and Justin Sydnor*

Financial incentives have shown strong positive short-run effects for problematic health behaviors that likely stem from time inconsistency. However, the effects often disappear once incentive programs end. This paper analyzes the results of a large-scale workplace field experiment to examine whether self-funded commitment contracts can improve the long-run effects of an incentive program. A four-week incentive program targeting use of the company gym generated only small lasting effects on behavior. Those that also offered a commitment contract at the end of the program, however, showed demand for commitment and significant long-run changes, detectable even several years after the incentive ended. (*JEL* D03, I10, J32)

Many people state a desire to change health-related behaviors, yet struggle to do so. These behavioral problems in domains such as weight loss, smoking, and exercising have helped to motivate a rich literature in economics on time-inconsistent behavior (Strotz 1955–1956; Laibson 1997; O'Donoghue and Rabin 1999, 2001; Loewenstein, O'Donoghue, and Rabin 2003; DellaVigna and Malmendier 2006). This literature has shown that time inconsistency can generate patterns of behavior that lead to both "internalities," where one's short-run actions are perceived as suboptimal from one's long-run perspective, and traditional externalities coming through higher group-rated health insurance costs and spending on Medicare and Medicaid (Finkelstein et al. 2009).

In the face of these problems, there is increasing interest from firms, insurance companies, policy makers, and health professionals in using financial incentives to motivate changes in health behaviors (Volpp et al. 2009a; Baicker, Cutler, and Song

*Royer: University of California, Santa Barbara, Mail Stop 9210, Santa Barbara, CA 93106 (e-mail: heather.royer@ucsb.edu); Stehr: Drexel University, 3220 Market Street, Philadelphia, PA 19104 (e-mail: stehr@drexel.edu); Sydnor: Wisconsin School of Business, University of Wisconsin, Madison, 975 University Ave., Madison, WI 53706 (e-mail: jsydnor@bus.wisc.edu). We are thankful for funding from the National Science Foundation grant numbers 0819804 and 1025846, the Upjohn Institute grant number 07-106-10, and the Case Western Reserve University ACES fund grant number 0245054. Royer also thanks the RAND Corporation for support through the NIA. We are appreciative for the outstanding research assistance of Sarah Bana, Stephen Cabrera, Andrew Chang, Vishal Cauhan, 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.

†Go to http://dx.doi.org/10.1257/app.20130327 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

2010). The use of incentives to promote health is particularly relevant for policy makers 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 but vibrant literature has emerged to explore the effect of incentive programs on changing health behaviors (Finkelstein et al. 2007; Volpp et al. 2008; Volpp et al. 2009b; Charness and Gneezy 2009; Acland and Levy 2015; Babcock and Hartman 2010; Babcock et al. 2011; Cawley and Price 2013; John et al. 2011; Just and Price 2013). While this literature has its limitations, overall it points to strong responses to financial 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, and Rey-Biel 2011; John et al. 2011). Understanding whether incentive programs can be designed to have stronger long-run effects is an important open question.

In this paper, we present the results of a large-scale randomized field experiment that explores whether self-funded deposit commitment contracts can be used as a low-cost method of improving the overall performance of a health-incentive program. Self-funded deposit commitment contracts allow users to put money at stake that is lost if the person does not meet a goal. This approach is motivated by the literature on time inconsistency, which highlights that individuals aware of self-control problems may value a way to incentivize their future behavior. In an experiment with smokers in the Philippines, Giné, Karlan, and Zinman (2010) show that this type of contract can be an effective tool for smoking cessation. These contracts are also available to the general public through the commercial website StickK.com, where over a quarter-of-a-million commitment contracts have been created. Our study provides the first large-scale experiment using deposit commitment contracts in the United States. Our primary innovation is to test whether the option of a self-funded commitment contract at the end of a health-incentive program improves the long-run effects of the program on behavior.

The field experiment involved 1,000 employees at a Fortune 500 company in the United States and was conducted over 2 years. Employees randomly selected for 1 of the 2 treatment groups were all offered a one-month financial incentive to attend their company's onsite exercise facility ($10 per visit for up to 3 visits each week). 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 option allowed participants to put money at stake for a pledge that they would continue to use the gym over the two 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.[1]

Employees responded strongly to the incentive program, with gym attendance doubling during the incentive period relative to the control group. While this increase is not in itself surprising, these treatment effects provide the first benchmark of incentive effects in a workplace setting from a large-scale randomized trial. From

[1] In other applications, such as StickK.com, commitment contracts are often given with the option for forfeited money to go to an undesirable source, what StickK.com calls an "anti-charity." The relative effectiveness of different forfeiture options remains an unstudied area.

supplemental survey data, we are able to establish whether this increase is due to new exercise or simply a change in the location of exercise. Our results suggest that 70 percent of the induced attendance is new exercise.

Our primary focus is the extent to which incentivized individuals continue to exercise at higher rates after the end of the incentive program (i.e., exhibit "habit formation"), and more specifically whether the availability of a commitment option improves its lasting impact.[2] Two recent studies of undergraduates find that similar exercise incentives lead to significant habit formation in the few months after the incentive program ends (Charness and Gneezy 2009; Acland and Levy 2015). In this study, we see some lasting changes for those who were not members of the corporate gym prior to the intervention. But overall the effects of the incentive program faded quickly. Those offered only the 4-week incentive retained only 25 percent of the incentive-induced increase in exercise frequency over the month following the intervention. Most of the increase disappeared within two months.

The availability of the commitment option at the end of the incentive program, however, significantly improved the lasting effect of the program, with no additional program cost. Of the incentivized employees offered the additional commitment option, 12 percent created a commitment contract, and the average commitment size was $58. Among those who had attended the gym at least once during the incentive period, the take up rate was 22 percent. Over the initial two months after the incentive ended (when the commitment contracts were in place), this group offered the commitment option retained half of their incentive-induced increase in exercise. During the commitment contract period, their gym attendance exceeded that of the control group by 50 percent and that of the incentive-only group by 25 percent.

The program combining incentives with a commitment option showed very long-run effects on behavior, even after the commitment-contract period had ended. For this group we see stable treatment effects of around 4 percentage points on the fraction of employees using the gym each week even 2 to 3 years after the incentive period ended, which represents a 20 percent increase relative to the control baseline.

We see two possible explanations for why the additional commitment option had stronger long-run effects extending after the end of the commitment-contract period. The first possibility is that once this group was introduced to the idea of commitment contracts, they continued to use some form of commitment device (e.g., Stickk.com) on their own to overcome motivational problems with their exercise after they no longer had the experimental commitment option. The second possibility is that the two month commitment-contract period allowed these employees to effectively lengthen their incentive period by incentivizing themselves. If the four-week incentive program was too short for some to establish a self-sustaining exercise habit, the ability to extend the incentive period could have helped generate stronger lasting changes in behavior. This second possibility highlights an important benefit of incorporating self-funded commitment options into incentive programs. It may be cost prohibitive or difficult to know the necessary length for an incentive

[2] Consistent with the existing literature we refer to these long-run effects as "habit formation" effects. However, we recognize that long-run effects may be due to channels such as learning about exercise that not all would refer to as "habit formation."

program to establish self-sustaining habits, especially since the optimal length may be different in different contexts. A commitment option may allow users to endogenously lengthen their incentive period to reach critical thresholds in habit formation without additional program costs.

This study also provides some new insights about the demand for commitment contracts. We find strong gender and body-weight gradients, with women and the overweight being substantially more likely to create a commitment contract. Most interestingly, the demand for commitment is present even among those who were exercising regularly prior to the study. Prior studies of commitment contracts of the type used here have targeted populations with clear behavioral issues (e.g., smokers, obese). As far as we know this is the first study that has explored the demand for commitment for those with no externally obvious behavioral problems. The fact that we observe demand for commitment by those who appear on the surface not to need it highlights that it can be difficult to identify those with time-inconsistency problems ex ante. One reason for this may be that individuals who are aware of time inconsistency problems find partial solutions to those problems in ways that mask behavioral problems and make it difficult to measure time inconsistency through surveys. In this case, financial commitment contracts may substitute for other forms of self-control, which in turn could mean that there are positive welfare effects of commitment technologies even if they do not generate large observable differences in behavior. Finally, although the literature on time inconsistency has tended to focus on the demand for commitment by "sophisticates" (O'Donoghue and Rabin 2001) who are perfectly aware of their level of present bias, we find some suggestive evidence that those who are somewhat (over)confident about their future behavior may be the ones with the highest demand for commitment.[3]

Our work adds to a small but growing literature on the use of commitment devices to overcome time-inconsistency. Giné, Karlan, and Zinman (2010), cited above, is the only other systematic study of the effectiveness of "pure" deposit commitment contracts, where deposited money can be lost but there is no financial upside to a commitment contract.[4] A few studies in the medical and public health literature targeting weight loss and smoking cessation have shown success using incentive programs with forms of deposit contracts with financial up-side benefits to the deposits (Jeffery, Hellerstedt, and Schmid 1990; Volpp et al. 2008; John et al. 2011). A number of studies also show that various commitment technologies that generally limit flexibility or access can be successful in promoting savings (Ashraf, Karlan, and Yin 2006; Benartzi and Thaler 2004; Beshears et al. 2011),[5] exercise (Milkman, Minson, and Volpp 2014), and work effort (Augenblick, Niederle, and Sprenger 2013; Kaur, Kremer, and Mullainathan forthcoming). Our results complement this emerging literature by showing the potential value of deposit commitments in a

[3] Bryan, Karlan, and Nelson (2010) argue that those with mild overoptimism may still see commitments as desirable but will likely believe that weak commitments be more effective than they are.

[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.

[5] Brune et al. (2013), however, observed that most used an uncommitted account over a commitment account when both were available. Karlan and Linden (2014) also find that a savings account with a full commitment to education was less effective than an account that was merely intended for educational expense.

large workplace experiment in the United States. We see that commitment contracts can be a promising new way of improving the lasting response of an incentive program for exercise at no additional program cost. Our results also indicate that commitment contracts might be used in conjunction with periodic incentives as a cost-effective alternative to permanent incentives. At a more general level, these results suggest that more research on the potential of commitment technologies for addressing challenges in maintaining behavioral change may be warranted.

The remainder of the paper is organized as follows. Section I describes the experimental design and provides summary statistics on the data. Section II presents the main results and Section III discusses substitution patterns away from other exercise. Section IV presents a range of analysis exploring heterogeneity and robustness of the effects. In Section V, we conclude with a discussion of some possible implications of this study and directions for future research.

I. Experimental Design and Data

A. *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 and an on-site fitness center. This gym offers showers, lockers, and towel service. The gym consists of two main rooms: one for classes such as yoga and aerobics and one with exercise equipment (e.g., treadmills, stationary bikes, and weights). 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.[6,7] 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. The cohorts were small enough that overcrowding was unlikely to occur; the increase in visits we show later resulted in an approximately 5 percent increase in total visits. The gym employees told us that the gym was operating under capacity. Our experiment had little impact on attendance during noon hour, the busiest time of the day, when one-third of the total daily gym visits happen.[8] Online Appendix Figure 1 describes the timeline of the experiment. We detail the number of participants along each step of the experiment in online 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

[6] There are no start-up fees or contracts and employees can cancel their membership at any time with no penalty.

[7] The gym is open Monday through Friday from 6:00 A.M. to 7:00 P.M.

[8] We observe a difference of less than one noontime visit during the time when our incentive program was in place.

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 five 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 percent (see online Appendix Figure 2).[9] We view this response rate as relatively high; for comparison, for the Card et al. (2012) study of peer pay of University of California employees, the survey response rate among employees was just above 20 percent. 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 1,000 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 treatment and control groups. The treatment group was eligible to receive financial incentives for gym attendance for a four-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 (i) whether the subject was an existing member of the company gym and (ii) whether they responded in the initial survey that their current exercise was at or above their personal exercise target. After the completion of the incentive period, the incentive-eligible subjects were divided into two treatment groups: incentive-only 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 percent (see online 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.[10] In light of the response e-mails we received, we suspect that a significant fraction of nonresponse was driven by those who traveled away from work during our recruitment. Of course, a company-sponsored program could be communicated to employees in a variety of ways, whereas our communication was mostly via e-mail. Since we feel

[9] 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 and this affected participation, we would expect that response rates and the fraction who are gym members would vary across cohorts.

[10] 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 nonmembers—74 percent versus 57 percent.

the observable characteristics we have for nonresponders 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 nonresponders would not respond to financial incentives. At the end of our experiment, we contacted nonresponders to our initial survey and assigned them to different treatment arms without requiring them 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 nonresponders would not change our conclusions qualitatively if extended to the full population.

B. *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 length of the intervention and the size of the incentives roughly followed that of Charness and Gneezy (2009).[11] (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.[12] To ensure that the incentives were salient to participants, we informed treatment subjects via both e-mail 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 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

[11] Charness and Gneezy (2009), however, do not use per-visit incentives and instead base incentives on whether individuals met or exceeded an eight visit threshold.

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

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 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 nonobese, which fits well with company-wide health promotion efforts. Fryer, Jr. (2011) 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 nonobtrusive 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 2013).

C. *Second-Level Treatment: Self-Funded Commitment Contract*

At the end of the four-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. Up until the commitment contract offer, we treated these groups the same. Throughout, incentive+commit denotes the treatment group *offered* the commitment option, and is an "intention-to-treat" grouping. Incentive-only refers to the treatment group that was not offered the commitment contract.[13] 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 toward 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 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

[13] In order to ensure balance between the incentive-only and the incentive+commit groups, we rerandomized 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 the first few cohorts. For that reason we decided to change the protocol and conduct the randomization after the incentive period for later cohorts.

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 eight-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 e-mail 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 e-mail to the incentive+commit group 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 e-mail 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.

D. *Data*

Table 1 provides the means for key variables from our initial survey.[14] 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 and 5) and between the control and incentive+commit group (columns 3 and 6).[15] The last two columns in each panel are the *p*-values from two tests: first, the equivalence of the means across the three randomized groups; and second, the equivalence

[14] 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. The size of each treatment group for each cohort was dictated by number of new possible members the wellness center could sign up. Treatment probabilities ranged between 0.2 and 0.4.

[15] 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.

Table 1A—Pre-Treatment Descriptive Statistics for Existing Members

	Control mean (1)	Incentive-only difference (2)	Incentive+commit difference (3)	p-value (2) = (3) = 0 (4)	p-value (2) = (3) (5)
Panel A. Demographics					
Age	40.12	−1.17	−0.01	0.61	0.37
Male	0.46	0.04	0.08	0.53	0.63
College degree or more	0.61	0.09	0.05	0.37	0.52
Panel B. Living situation					
Married	0.68	0.02	0.04	0.78	0.71
Has at least 1 kid at home	0.45	0.08	0.07	0.46	0.89
One-way commute (minutes)	37.82	−2.35	−0.21	0.48	0.28
Panel C. Subjective well-being					
Unhappy with life	0.07	0.00	0.02	0.89	0.70
Unhappy with fitness	0.34	0.05	0.01	0.64	0.53
Unhappy with weight	0.58	−0.01	−0.08	0.33	0.23
Panel D. Health and fitness					
Pounds over target weight	20.28	−1.63	1.42	0.69	0.37
BMI	28.31	−0.60	−0.29	0.68	0.67
Overweight	0.43	−0.04	0.03	0.50	0.29
Obese	0.30	0.00	−0.06	0.47	0.30
Takes blood pressure meds	0.12	0.03	0.00	0.78	0.57
Panel E. Exercise					
Average days of overall exercise	3.36	−0.10	0.12	0.34	0.16
Target days of exercise	4.79	0.03	0.19	0.26	0.17
0 days of overall exercise	0.05	0.03	−0.03	0.16	0.07
Observations	94	134	131		

Notes: Column 1 is the control group mean. Columns 2 and 3 are the estimated mean differences between that group and the control group. Estimates for columns 2 and 3 come from regressions with strata fixed effects.

Source: Authors' calculations

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 percent level.

Our subject pool is on average 40 years old, roughly equally divided across genders, and is well-educated (more than 65 percent have a college degree or more). In comparison, overall in the United States in 2009, just under 30 percent 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.[16] Company employees are on average somewhat less unhappy than in the United States as a whole (14.3 percent report being unhappy in the 2010 General Social Survey).[17]

[16] Baseline statistics for this and previous sentence based on authors' calculations using the 2010 census.
[17] Source of statistic is http://sda.berkeley.edu/cgi-bin/hsda?harcsda+gss10.

Table 1B—Pretreatment Descriptive Statistics for Nonmembers PreStudy

	Control mean (1)	Incentive-only difference (2)	Incentive+commit difference (3)	p-value (2) = (3) = 0 (4)	p-value (2) = (3) (5)
Panel A. Demographics					
Age	39.62	−0.75	0.16	0.65	0.36
Male	0.52	0.01	−0.01	0.90	0.71
College degree or more	0.64	0.03	0.10	0.06	0.09
Panel B. Living situation					
Married	0.67	−0.03	0.01	0.65	0.31
Has at least 1 kid at home	0.48	−0.05	0.04	0.19	0.06
One-way commute (minutes)	38.03	0.76	−0.85	0.69	0.41
Panel C. Subjective well-being					
Unhappy with life	0.11	0.01	0.01	0.91	0.86
Unhappy with fitness	0.54	−0.10	−0.08	0.08	0.72
Unhappy with weight	0.55	−0.04	−0.05	0.61	0.90
Panel D. Health and fitness					
Pounds over target weight	22.72	−0.91	2.00	0.58	0.29
BMI	28.22	−0.49	0.40	0.36	0.13
Overweight	0.42	−0.06	−0.04	0.44	0.62
Obese	0.30	−0.02	0.03	0.43	0.18
Takes blood pressure meds	0.13	0.00	−0.02	0.75	0.49
Panel E. Exercise					
Average days of overall exercise	1.98	−0.13	−0.09	0.54	0.73
Target days of exercise	4.05	−0.18	0.00	0.30	0.20
0 days of overall exercise	0.24	0.01	0.02	0.86	0.76
Observations	195	228	215		

Notes: Column 1 is the control group mean. Columns 2 and 3 are the estimated mean differences between that group and the control group. Estimates for columns 2 and 3 come from regressions with strata fixed effects.

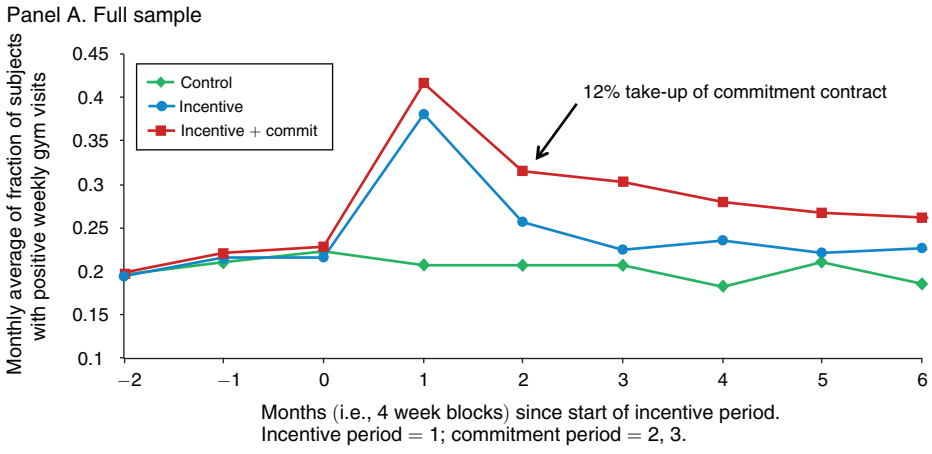
Source: Authors' calculations

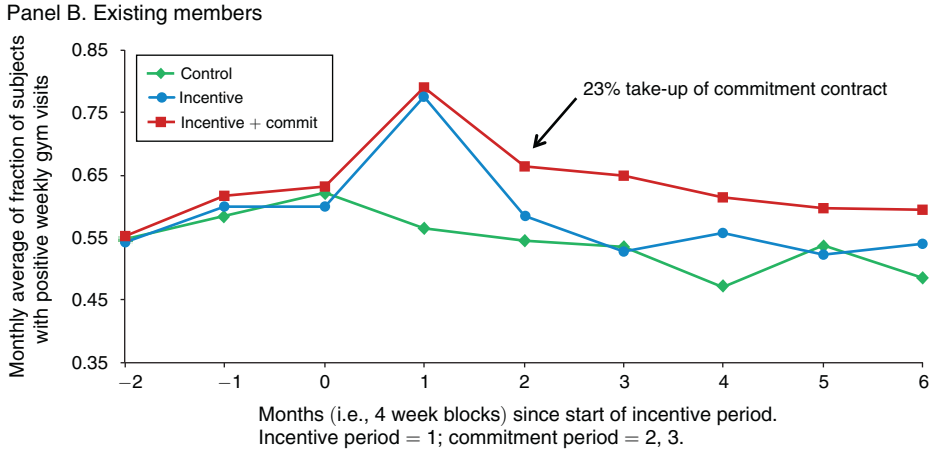
We asked subjects in the initial survey to report their current exercise activities and their targets for how often they would like to exercise. The average difference between targeted and self-reported exercise is 1.5 days/week for gym members and 2 days/week for nongym members, implying that individuals 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 fraction of individuals reporting no exercise in a typical week. Thus, our subjects likely have much to gain from increased exercise.

II. Results

A. *Graphical Comparisons of Means*

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 4-week average of the fraction attending the gym at least once in the week (e.g., for an individual who does not attend in weeks one and two





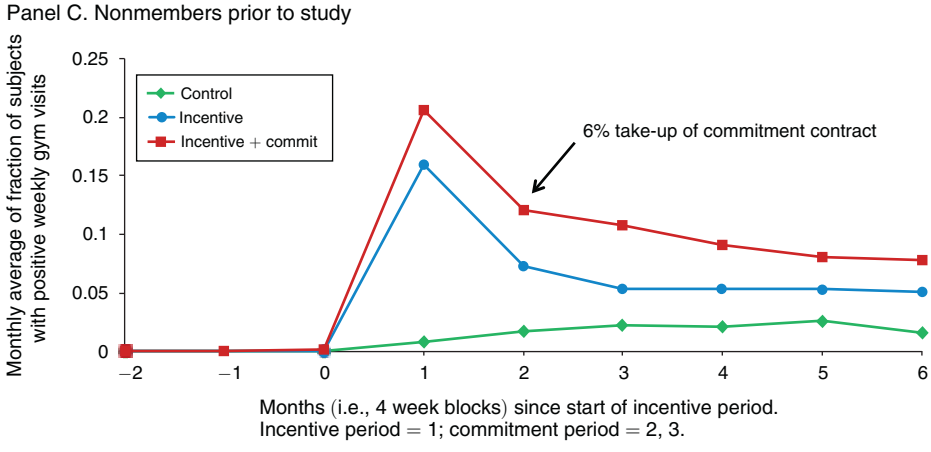


FIGURE 1. FRACTION WITH POSITIVE GYM VISITS BY TREATMENT

and attends 3 times in weeks three and four, this average would be 0.5). We combine the data for each cohort such that month one is the four-week incentive period. Months two and three encompass the period of the commitment contract.[18] The graphs go out for six months from the beginning of the treatment period.

Figure 1 panel A shows the overall results. As we would expect from random assignment, all three groups (control, incentive-only, and incentive+commit) had similar pretreatment patterns, with on average approximately 20 percent 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 posttreatment patterns. The group offered only incentives largely reverts toward baseline behavior but with a small lasting increase in visit frequency relative to control. In contrast, the attendance probability of the incentive+commit group, 12 percent of whom decided to create a commitment contract, remains clearly elevated relative to both pretreatment levels and the control group over time. During months two and three, when the commitment contracts were in place, approximately 30 percent of the incentive+commit group attended the gym at least once per week, while the control remained at the 20 percent baseline and the incentive-only group fell from around 25 percent in month two to just over 20 percent in month three.

After the end of the commitment-contract period, the attendance rates of the incentive+commit group fall somewhat but remain clearly elevated relative to control. Since the commitment-contracts were no longer in place after month three, the lasting effect of the incentive+commitment treatment is particularly striking. It is difficult to know exactly what mechanisms underlie this elevated effect relative to what we see for the group offered only incentives. 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. Another possibility is 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 1, panels B and C present time series separately based on gym-membership status prior to the experiment, a stratification variable for randomization. Figure 1,

[18] 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.

panel B 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 percent of existing members using the gym at least once in an average week.[19] That fraction rose to 80 percent 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 three, and shows no real lasting response to the incentive. In contrast, the incentive+commit group (23 percent commitment take-up) shows a lasting response to the incentive program. Their attendance rates are approximately 10 percentage points higher than the control during months two and three and remain clearly elevated at the six-month mark. Figure 1c shows the patterns for those who were not members prior to the experiment.

Overall the incentive program motivated 18.2 percent 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 six months after the end of the incentive program. This suggests that for a modest number of employees the temporary incentive program generated a lasting shift in the use of the company gym. The postincentive program effects for the incentive+commit group are even higher relative to control. The incentive+commit group attendance exceeds that of the incentive-only group for the entire postincentive period. 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.

B. *Regression Framework*

To quantify our results and establish confidence intervals on the effects we run simple OLS regressions. Our regression models take the following form:

$$(1)\qquad y_{itw} = \tau_0 + \alpha_1 T_1 + \alpha_2 T_2 + \sum_{j=1}^{n}\left[\gamma_j p_j + \beta_{j,1} p_j T_1 + \beta_{j,2} p_j T_2\right] + \mu_s + \pi_w + \varepsilon_{itw},$$

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. The treatments are indicated by T_1 for the incentive-only group, and T_2 for the incentive+commit

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

group. In any specification we have indicator variables for time period (p_j) (e.g., in-treatment incentive period) and the omitted time period is the six-week period prior to when we first contacted any employees about the study. With this structure the γ_j coefficients measure how the control group's use of the gym evolves across periods as compared to the pretreatment period. The coefficients α_1 and α_2 provide an estimate of how the incentive and incentive+commit mean outcomes differed from that of the control in the pretreatment period and should be near zero due to randomization. The coefficients on the interactions between the treatment indicators and the periods ($\beta_{j,1}$ and $\beta_{j,2}$) are our primary outcomes of interest and can be interpreted as difference-in-differences parameters. $\beta_{j,1}$ and $\beta_{j,2}$ measure the extent to which changes in the mean outcomes between the intervention and the preintervention periods for the incentive only and incentive+commit group, respectively, differ from the analogous change for the control group. We account for the strata used in randomization (the combination of exercise versus target, ex ante company gym membership status, and cohort) with fixed effects denoted here by μ_s. We also include π_w, which are calendar week fixed effects estimated separately for members and nonmembers as a means of variance reduction since the outcomes have a large seasonal component. Since there are weekly observations on the same individuals, we adjust the standard errors for clustering at the individual level.

C. *Regression Results for In-Treatment and Early Posttreatment Effects*

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 nonmembers (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.[20] 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 0 otherwise. In columns 2, 4, and 6, the dependent variable is 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 preperiod 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 percent 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

[20] 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.

Table 2—Ordinary Least Squares Regressions of Treatment Effects

	Overall		Members		Nonmembers	
	Any visit (1)	Weekly visits (2)	Any visit (3)	Weekly visits (4)	Any visit (5)	Weekly visits (6)
Mean for control group in preperiod:	0.20	0.58	0.62	1.80	NA	NA
Incentive only	−0.01	−0.07	−0.02	−0.19		
	(0.02)	(0.06)	(0.05)	(0.16)		
Incentive+commit	0.00	−0.00	0.01	−0.03		
	(0.02)	(0.06)	(0.05)	(0.16)		
In-treatment period (weeks 1–4)	0.02	0.06	0.00	0.01	0.03***	0.07***
	(0.01)	(0.04)	(0.03)	(0.12)	(0.01)	(0.03)
(Incentive only) × (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) × (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 posttreatment (weeks 5–13)	0.03**	0.09**	0.00	0.03	0.05***	0.12***
	(0.01)	(0.05)	(0.03)	(0.11)	(0.01)	(0.04)
(Incentive only) × (early post)	0.04**	0.12**	0.03	0.15	0.04**	0.09*
	(0.02)	(0.05)	(0.03)	(0.11)	(0.02)	(0.05)
(Incentive+commit) × (early post)	0.09***	0.24***	0.10***	0.30***	0.09***	0.21***
	(0.02)	(0.05)	(0.03)	(0.10)	(0.02)	(0.06)
Late posttreatment (weeks 14–26)	0.04**	0.14**	0.01	0.09	0.06***	0.16***
	(0.02)	(0.06)	(0.05)	(0.15)	(0.02)	(0.06)
(Incentive only) × (late post)	0.04**	0.11**	0.04	0.14	0.03**	0.09*
	(0.02)	(0.05)	(0.04)	(0.12)	(0.02)	(0.05)
(Incentive+commit) × (late post)	0.07***	0.21***	0.08**	0.25**	0.06***	0.18***
	(0.02)	(0.06)	(0.04)	(0.12)	(0.02)	(0.06)
Subject-week observations	32,000	32,000	11,488	11,488	20,512	20,512
Number of subjects	1,000	1,000	359	359	641	641
p-values test of equal effects—incentive-only versus incentive+commit:						
Pretreatment	0.35	0.20	0.46	0.25	0.25	0.25
Intreatment (weeks 1–4)	0.35	0.12	0.72	0.53	0.16	0.13
Early posttreatment (weeks 5–13)	0.002	0.03	0.03	0.16	0.02	0.08
Late posttreatment (weeks 14–26)	0.07	0.07	0.21	0.27	0.27	0.15
Commitment contract take-up:	0.12	0.12	0.23	0.23	0.06	0.06
IV estimate week 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 by exercise above or below target by member (in pooled regressions) fixed effects) and separate week fixed effects. For the overall regressions, separate week fixed effects for members and nonmembers are included. The IV estimates are measured as the effect of a commitment contract on gym attendance using assignment to 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., the number of visits in week 1 of the incentive period for the visits dependent variable) and strata fixed effects.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Source: Authors' calculations

the table we display *p*-values from tests of the equivalence of the incentive-only and the incentive+commit group coefficients in the preincentive, incentive, and early and late postincentive periods. Since the groups were treated the same during the incentive period, we expect the in-treatment results to be similar. The *p*-values confirm this expectation on the equivalence of incentive period effects for the incentive and incentive+commit groups.

The estimates from the early posttreatment 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 (months 2 and 3 in Figure 1). Consistent with the graphical results, we find that during the first two months after the incentive program, the probability of attending the gym is slightly elevated (0.04) 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 22 percent (0.04/0.18) of their increase in visit frequency relative to the control over these two months. The probability of attendance for the incentive+commit group was 9 percentage points higher than the control over this period; the commitment period effect is 45 percent of the in-treatment effect. The effects for the incentive+commit group in the early posttreatment 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.

One reasonable question is whether these effect sizes for the incentive+commit group in the early posttreatment period 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 percent overall. Not surprisingly, the commitment rate of members exceeded that of nonmembers. However, when excluding those who did not attend the gym during the incentive period, the commitment rates are similar: 23 percent for members and 21 percent for nonmembers. 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 above and beyond the incentive effect for the early posttreatment 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 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.[21] In other words, the size of these effects is consistent with the idea that the long-run effects on gym attendance among the incentive+commit group are due to the commitment itself.

The estimates for the late posttreatment period (weeks 14–26) are in line with the effects we see for months 5 and 6 in Figure 1. The increases relative to control in visit frequency for the incentive-only group are very similar to those estimated for

[21] 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.

the early posttreatment period. The incentive+commit group shows a slight reduction of their effect in this period after the commitment contract option ended but shows a significant effect size relative to the control group. We also see marginally statistically significant differences between the incentive+commit and incentive-only effect sizes (p-values of 0.07 in both columns 1 and 2) for this period. The combination of incentives with the commitment contract option appears to have had a lasting effect even after the commitment contracts ended.

We can compare the incentive effect sizes to two recent studies with undergraduate populations, Charness and Gneezy (2009) and Acland and Levy (2015), 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 (2015) 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 posttreatment effect for Charness and Gneezy (2009) is 0.59, whereas for Acland and Levy (2015) it is 0.26. Our estimate of initial posttreatment effects for employees offered only incentives is again substantially smaller at 0.12. Expressed as a ratio of the in-treatment effect, the observed posttreatment effects in our study for the incentive-only group are close to those in Acland and Levy (2015) and about half the size observed by Charness and Gneezy (2009).

D. *Heterogeneity by Ex Ante Gym Membership Status*

In columns 3–6, 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 early posttreatment 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 small differences in any visit patterns for those who received incentives only. We find modest long-run effects for members in the incentive+commit group, consistent with the estimates for the pooled sample, but these differences are not statistically significant with the reduced sample size of members only.

For nonmembers in the incentive-only group, we estimate statistically significant increases in gym attendance in both the early and late posttreatment 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 nonmembers to gym users relative to the control. Compared to the in-treatment effects, around 25 percent of the new gym use effect due to the incentives for this group is retained in the long run. The difference between the incentive only and the incentive+commit group is statistically-significant for the early posttreatment period but not for the late posttreatment period. Nonmembers in the incentive+commitment group had a 9 percentage point increase in the fraction attending the gym relative to control in the early posttreatment period, which declines to 6 percentage points in the late posttreatment period.

Of course, for this nonmember population, one concern with the comparison of behavior for those offered incentives only versus those also offered commitments is the differential in-treatment response (albeit not statistically significantly different from one another) to the incentive program between these two groups. To address such concerns, we run separate analyses where we control for in-treatment visits through either the inclusion of a set of four dummy variables indicating whether an individual attended the gym in a particular week of the incentive period or a set of fixed effects for visits for each week of the incentive period. We consistently find differences in the usage patterns between incentive-only and incentive+commit groups during the early posttreatment period. Online Appendix Table 2 displays the estimated differences between the incentive+commit and incentive-only groups (the control group is excluded). For nonmembers, our estimate of the early posttreatment difference between the incentive+commit and incentive-only groups 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-only and incentive+commit group have little impact on our conclusions about the long-run effectiveness of the commitment contract for nonmembers.

E. *Long-Run Effects*

One of the benefits of our study is that we have access to records of gym attendance for these employees over a long time horizon and can estimate very long-run effects of the treatments. This is an important advance over previous studies of health incentives, whose long-run follow up has been limited by short-run access to clinical populations, substantial attrition in follow-up measurements, or short-run follow up due to end of the school term in studies with undergraduate populations.

We begin our analysis of long-run effects with Figure 2, which shows the evolution of the treatment effects. Here we graph the estimated treatment effects looking at the outcome of fraction of employees with at least one visit per week over eight-week periods running through three years after the start of the treatment program.[22]

The 0-mark on the x-axis shows the estimated effect during the treatment period, while weeks 8–13 are denoted as months 1–2, which was also the period of the commitment contracts. The remaining two-month points denote subsequent eight-week blocks. We show the 95 percent confidence intervals around each estimated coefficient based on standard errors that are clustered at the individual level.

Figure 2, panel A shows the effects for the group offered incentives only relative to the control group. The estimated treatment effect of 3–4 percentage points documented before holds through the first year after the start of incentives, though none of the individual period estimates are statistically significant after month two. For this group the point estimates suggest that the treatment effects fall to zero about two years after the initial incentive period.

[22] The estimates come from a regression specification following the format described in Section IVB. Period 1 in the regression is weeks 1–4, period 2 is weeks 5–13 (commitment-contract period), and then periods 3 and on are the subsequent 8 week groupings.

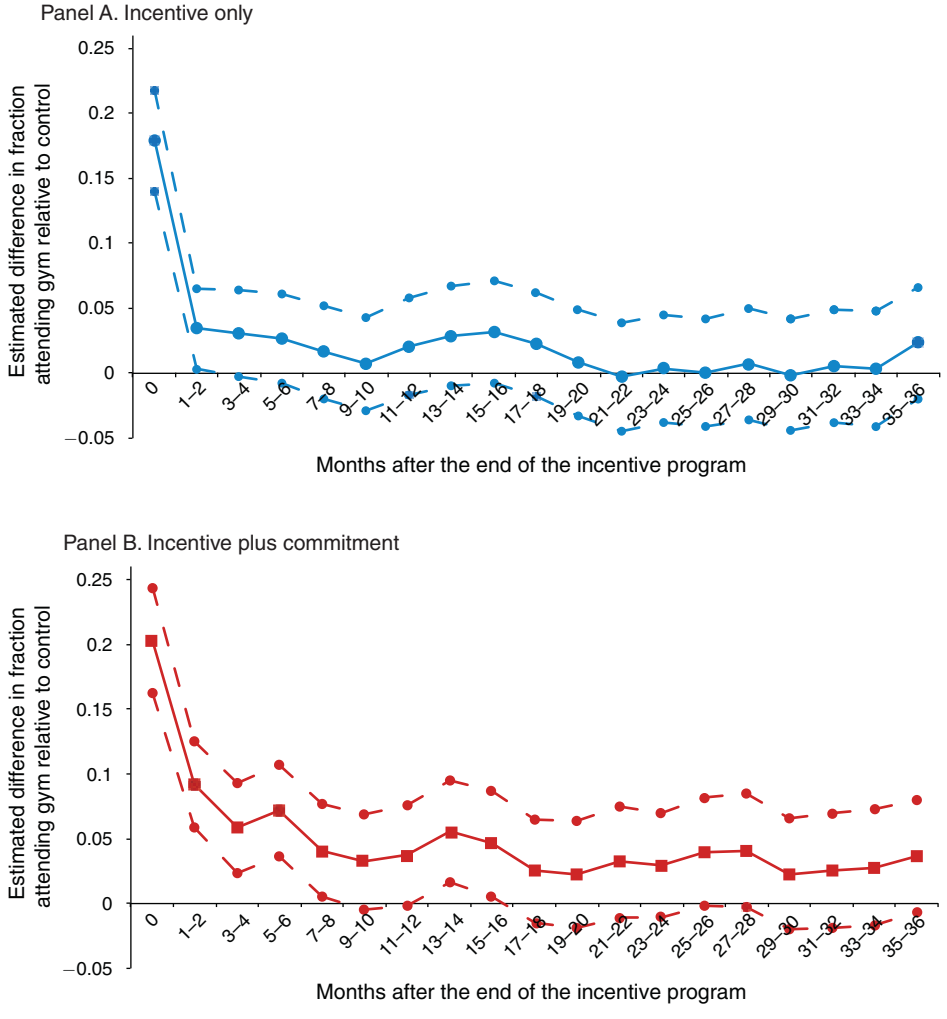


FIGURE 2. LONG-RUN TREATMENT EFFECTS RELATIVE TO CONTROL

Notes: Figure 2, panels A and B graph the estimated difference in the fraction of subjects attending the company gym at least once per week for different time periods. The estimates are obtained from a regression of an indicator for attending the gym in a given week on period dummies and period dummies interacted with treatment status. The x-axis value of 0 shows the estimated treatment differences for the four-week incentive period. The subsequent x-axis marks show the estimated effect from the interaction of the treatment status and an indicator for each two-month period (eight weeks) following the end of the incentive program. The regressions also control for experimental cohort and strata fixed effects, calendar-week fixed effects, and for whether the individual was a member of the company gym prior to the experiment. Ninety-five percent confidence intervals around the various two-month effect estimates are given in dashed lines, with standard errors clustered at the individual level.

Figure 2, panel B shows the long-run effects for the group offered incentives plus the commitment option. Here we see that after the first 6 months the effects stabilize at around 5 percentage points over the remainder of the first year. The individual period effects are statistically significant at either the 5 percent or 10 percent level through the sixteenth month. Over the next two years after that the treatment effect estimate stabilizes at around 3–4 percentage points, though none of these very long-run effects estimated separately by period are statistically significant. This

Table 3—Ordinary Least Squares Regressions of Posttreatment Effects

	Overall		Members		Nonmembers	
	Any visit (1)	Weekly visits (2)	Any visit (3)	Weekly visits (4)	Any visit (5)	Weekly visits (6)
Mean for control group in preperiod:	0.20	0.58	0.62	1.80	NA	NA
Later posttreatment (weeks 27–52)	−0.01	−0.02	−0.04	−0.13	0.01	0.04
	(0.01)	(0.05)	(0.04)	(0.13)	(0.01)	(0.03)
(Incentive only) × (later post)	0.02	0.05	−0.01	−0.02	0.03**	0.09**
	(0.02)	(0.06)	(0.04)	(0.14)	(0.01)	(0.04)
(Incentive+commit) × (later post)	0.04***	0.12**	0.03	0.09	0.05***	0.15***
	(0.02)	(0.06)	(0.04)	(0.13)	(0.02)	(0.05)
Latest posttreatment (weeks 53–104)	−0.01	−0.00	−0.06	−0.17	0.02*	0.08**
	(0.02)	(0.05)	(0.05)	(0.15)	(0.01)	(0.03)
(Incentive only) × (latest post)	0.02	0.04	0.02	0.04	0.02	0.05
	(0.02)	(0.06)	(0.04)	(0.15)	(0.02)	(0.04)
(Incentive+commit) × (latest post)	0.04**	0.12*	0.05	0.16	0.04**	0.10*
	(0.02)	(0.06)	(0.04)	(0.15)	(0.02)	(0.05)
Subject-week observations	110,000	110,000	39,490	39,490	70,510	70,510
Number of subjects	1,000	1,000	359	359	641	641
p-values test of equal effects—incentive-only versus incentive+commit:						
Weeks 27–52 posttreatment	0.11	0.18	0.20	0.38	0.28	0.26
Weeks 53–104 posttreatment	0.27	0.19	0.34	0.30	0.51	0.38

Notes: Robust standard errors clustered by individual in parentheses. All regressions include strata fixed effects (i.e., study cohort by exercise above or below target by member (in pooled regressions) fixed effects) and separate week fixed effects. For the overall regressions, separate week fixed effects for members and nonmembers are included. All regressions also include treatment dummies for the incentive and incentive+commit groups as well as treatment-period dummies for the in-treatment, early posttreatment, and late posttreatment periods from Table 2. In addition, the regressions include interactions between the treatment dummies and these treatment-period dummies.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Source: Authors' calculations

figure highlights that the combination of a one-month financial incentive and a two-month commitment option appears to have led to a permanent change in attendance behavior, with stable treatment effects of around 4 percentage points (20 percent increase) going out 3 years from the beginning of the incentive period.

In Table 3, to increase power in this long-run analysis, we present regression estimates analogous to those in Table 2 that explore the average treatment effect over long-run horizons of the remainder of the first year (weeks 27–52) and the second year after the study began (weeks 53–104). The estimates are consistent with the graphical analysis in Figure 2. In particular, we see stable treatment effects for the incentive+commit group of 4 percentage points in the fraction visiting the gym that are statistically significant for the first and second years after the study.[23] Effect sizes for incentive+commit treatment are quite similar among members and nonmembers although only for the nonmembers are the effects statistically significant. The magnitude of the impacts for the latest posttreatment period is over half

[23] We have also estimated the average treatment effect over the third year on the fraction attending, which is 0.032 with a *p*-value of 0.105.

the magnitude of that observed during the week 14–26 period and about 20 percent the size of the incentive period effects.

F. *Commitment-Contract Take-Up*

Overall among the 346 subjects in the incentive+commit group, 12 percent (43 subjects) chose to make a commitment. Among ex ante gym members, the take-up rate of commitments was 23 percent. For those who were not members of the gym prior to the study, the overall take-up rate was 6 pecent, but take-up was 21 percent 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, Giné, Karlan, and Zinman (2010) saw an 11 percent take-up of their smoking cessation commitment device in the Philippines. Ashraf, Karlan, and Yin (2006) had a 28 percent take-up of their commitment savings product. The average commitment contract size was $58 and the maximum size was $300.[24] Comparing the size of commitment contracts to the money subjects earned during the incentive program, we see that 40 percent of the commitments match the incentive earnings, while 49 percent committed something less than their full incentive earnings and 11 percent committed an amount greater than their incentive earnings. 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 4 we present regression results examining the correlates of commitment-contract demand. For this analysis, we restrict the sample to subjects in the incentive+commit group who 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.[25] This leaves us with a sample of 194 subjects for this analysis (56 percent of the incentive+commit group). 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 22 percent. We need to be somewhat cautious in interpreting these regressions, since they rely on nonexperimental variation and were not generally prespecified. However, we see even an exploratory exercise as potentially valuable given that the few large field experiments on deposit commitment contracts.

The first column of Table 4 investigates how prior exercise experience relates to the demand for commitment. The top row of the table controls for the total number of visits made during the treatment period, which allows us to account for earnings through the incentive period and "house money" effects.

The next set of variables in the table are of greater interest and control for categories of exercise behavior prior to our study. We can break our sample up into four conceptually distinct groups: members versus nonmembers crossed with a split by high and low exercise frequency prior to the study. Among members, who all pay

[24] 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.

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

Table 4—Ordinary Least Squares Regressions Predicting Take-Up of Commitment Option

	(1)	(2)	(3)	(4)	(5)
Mean take-up rate:	0.22	0.22	0.22	0.22	0.22
Total visits during the treatment period	−0.01	−0.002	−0.002	−0.002	−0.003
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Prestudy exercise frequency (*omitted category: low-use members*)					
Members above median	0.21**	0.20**	0.23***	0.20**	0.19**
	(0.09)	(0.08)	(0.09)	(0.08)	(0.08)
Nonmembers above median	0.12	0.14	0.15	0.14	0.11
	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)
Nonmembers below median	−0.01	−0.02	−0.03	−0.03	−0.003
	(0.08)	(0.07)	(0.07)	(0.08)	(0.07)
Demographics					
Male		0.25***	0.24***	0.25***	0.25***
		(0.06)	(0.06)	(0.06)	(0.06)
College degree		0.04	0.04	0.04	0.03
		(0.06)	(0.06)	(0.06)	(0.06)
Overweight or obese		0.18***	0.18***	0.18***	0.17***
		(0.06)	(0.06)	(0.06)	(0.06)
Measures of awareness of self-control problem in presurvey					
States exercise is below personal target			0.06		
			(0.07)		
Rates self-control for exercise as low				0.01	
				(0.06)	
Chance of hitting personal exercise target next month					
Quartile 1 [0–50%)					−0.03
					(0.09)
Quartile 2 [50 –70%]					0.04
					(0.09)
Quartile 3 (70–85%]					0.09
					(0.10)
Number of subjects	194	194	194	193	194
R^2	0.04	0.14	0.14	0.13	0.15

Notes: Dependent variable is indicator for take-up of commitment contract. Robust standard errors in parentheses. Regression limited to those in the incentive+commit group (offered commitment option) who also stated they had some level of interest in attending the company gym in the follow-up survey conducted during week four of the study period (prior to anyone learning about the commitment option). All variables relating to expectations of behavior were asked in surveys prior to subjects learning about treatments and hence should measure subjects' ex ante beliefs about their likely behavior in the absence of treatment. The categories for prestudy exercise frequency are based on median splits separately for the groups of existing members and nonmembers.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Source: Authors' calculations

a monthly membership fee ($26), we can split by the median number of weeks in the 6 week prestudy period that they attended the gym (0–4 versus 5–6). This split is interesting because the frequent users by this definition were attending the gym often enough prior to the study to not be near the margin of the commitment contract, which required a visit only at least once every two weeks. In contrast, the low-frequency members are essentially the group of individuals identified by

DellaVigna and Malmendier (2006) who hold costly gym memberships without using them. This is likely the population that could benefit from commitment, but whether they are self-aware enough of the potential value of commitment to benefit from the option is an open question. We also include controls for nonmembers split by the median number of self-reported days of typical exercise in the initial study survey (1.5 days per week).

The key, and surprising finding, in column 1, is that the members with high prestudy visit frequency were actually substantially more likely than members with infrequent visits to make commitment contracts (21 percentage points). This effect is quite stable across specifications that add additional controls. Conditional on the number of visits during the treatment, we do not detect a statistically significant difference in commitment rate between low-use members and the two nonmember groups. However, for the nonmembers we again see that the point estimates suggest that those with higher levels of exercise are if anything more likely to make commitments. These findings suggest that commitment contracts may be at least as attractive for those who already have successful exercise routines as they are for those who appear more obviously in need of a commitment device.

Column 2 introduces demographic controls from the preintervention survey: gender, college degree, and overweight/obesity. Men are significantly less likely (25 percentage points) to make commitment contracts than women. We also find that being overweight or obese (as determined by self-reported weight and height) significantly positively predicts a takeup of commitment (18 percentage points). Finally, we do not find any significant correlation between commitment takeup and having a college degree. Given that college degree is likely the best proxy we have for socioeconomic status and income, the lack of correlation is at least suggestive that within this population commitment demand is not highly income dependent.

The remaining columns in the table explore the ability of proxies for awareness of a self-control problem to predict commitment demand. The behavioral literature suggests that the demand for commitment stems from those who are *aware* of their own time inconsistency problems (O'Donoghue and Rabin 1999, 2001). Unfortunately, however, it is very difficult to measure time inconsistency and even more difficult to derive measures of awareness of time inconsistency. As Augenblick, Niederle, and Sprenger (2013) discuss and demonstrate in experiments, the typical approach to identifying dynamically inconsistent time preferences using time-dated monetary rewards has significant confounds. While Ashraf, Karlan, and Yin (2006) used that method and found their measure of time inconsistency correlated with commitment demand for women (though not for men), Brune et al. (2013) find no correlation between measures of hyperbolic discounting and demand for commitment savings in their study. In our own pilot testing we found that similar questions generated almost zero instances of measured time inconsistency in our population of working US adults.[26] Augenblick, Niederle, and Sprenger (2013) show that real-effort tasks can be used to measure time inconsistency in a way that correlates with commitment

[26] We see this as primarily a result of the difficulty of assessing time inconsistency using a series of hypothetical questions related to time-dated monetary receipts.

demand, but there are not yet real-effort tasks of this type that can be feasibly adapted to a setting like ours.[27]

Recognizing the limitations in our ability to measure awareness of time inconsistency, in columns 3–5 of Table 4, we include variables collected through our surveys that may be partial proxies for an awareness of a time inconsistency problem. In column 3,, we include an indicator for individuals who stated in our prestudy survey that their level of exercise was below their personal target for exercise. Although this type of measure is imperfect, this seems like the most natural baseline measure for whether a person is aware of a time-inconsistency problem affecting their exercise.[28] We estimate a 0.06 positive association between being below the personal exercise target at commitment take-up, though the effect is not statistically significant. This relatively modest difference in takeup rate of commitments between those who say they are exercising less than they desire and those who say they are at their target is again consistent with the idea that those who appear to have successful exercise routines nonetheless demand commitment. In column 4, we include a dummy for whether the subject rated their own self-control for exercise as "low" in our presurvey and we find essentially no correlation between this measure and commitment takeup.

In the last column of the table we add controls related to the subjects confidence in their future behavior. We asked subjects in our presurvey to state what they thought the chance was that they would exercise up to their personal exercise target over the next four weeks.[29] Interestingly, we find that among the control group that almost none of those who were below their target and stated a likelihood greater than 50 percent of reaching their target actually did so over the following four weeks. This suggests that to some extent those expressing confidence in an ability to reach their goals are likely overconfident. In column 5, we include controls for the quartile split on their answer to the chance question. We find that relative to those in the top quartile (90–100 percent), those in the middle 2 quartiles were somewhat more likely to make commitments, while those in the bottom quartile (0–50 percent) were somewhat less likely. Although none of those effects are statistically significant, they are at least directionally suggestive that demand for commitment is highest among those with partial confidence in their ability to achieve their goals. The link between confidence and commitment takeup could come in part from individuals who are overconfident in the value of the modest commitment. It could also come in part from those who were ex ante confident because they intended to use other strategies to overcome their self-control problems and then substituted into the financial commitment contract when it was available.

[27] Giné et al. (2012) present another method for identifying time inconsistency through reversals in previous time-dated monetary allocations. Consistent with the discussion in Augenblick, Niederle, and Sprenger (2014), they find preference reversals but not strong indication of present-biased time inconsistency.

[28] There are at least two possible problems with this measure. First, "personal targets" reported in surveys may not conform to the theoretically relevant construct of the frequency of exercise a person would do in the absence of time-inconsistency problems. Second, those who have found other sources to overcome a time inconsistency problem already may report being in line with their personal target yet still have an underlying time inconsistency problem.

[29] This question was asked prior to subjects learning about the incentive program.

Taken all together, our analysis of commitment demand suggests two interesting findings. First, and most clearly, there are gender and body weight divides, with women and the overweight making commitments at much higher rates all else equal. Second, there is surprisingly little correlation between patterns of behavior (i.e., low exercise frequency prior to the study) or survey questions that suggest a likely need to solve a time inconsistency problem and the takeup of commitment. If anything, those who appear less in need of commitment may be more likely to use the commitment contracts. One potential explanation for both our findings on prestudy exercise patterns and the lack of predictive power of survey measures is that there may be significant populations who have time inconsistency problems that find ways to overcome them. Those people may then show no patterns of problematic behavior and yet still find financial commitment contracts to be a desirable substitute or complement to their prior methods of exerting self control. We return to this issue in our discussion in the concluding section.

III. 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.[30] 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.

Table 5 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 prespecified before the start of the experiment. We provide estimates for members and nonmembers separately. Panel A in the table shows results for all the subjects in each of those subgroups. The remaining panels delve further into possible heterogeneity, dividing the sample by whether an individual's preintervention 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

[30] 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 5—Substitution Analysis Comparing Incentive Effects in Gym Data to Survey Data

	Members			Nonmembers		
	Weekly visits	Weekly visits	Overall exercise	Weekly visits	Weekly visits	Overall exercise
Data source:	Gym (1)	Survey (2)	Survey (3)	Gym (4)	Survey (5)	Survey (6)
Panel A. All subjects						
Treated with incentive	0.79***	0.49***	0.40**	0.45***	0.60***	0.38***
	(0.15)	(0.17)	(0.18)	(0.05)	(0.07)	(0.13)
Observations	359	335	337	641	571	572
Mean control	1.59	2.26	3.25	0.03	0.07	2.09
Panel B. Subjects reporting exercise below their target in presurvey						
Treated with incentive	0.88***	0.76***	0.76***	0.47***	0.65***	0.45***
	(0.18)	(0.21)	(0.22)	(0.06)	(0.08)	(0.15)
Observations	209	190	192	499	446	447
Mean control	0.97	1.46	2.32	0.003	0.03	1.58
Panel C. Subjects reporting exercise at/above their target in presurvey						
Treated with incentive	0.60**	0.03	−0.16	0.37***	0.43**	0.11
	(0.26)	(0.29)	(0.29)	(0.14)	(0.19)	(0.31)
Observations	150	145	145	142	125	125
Mean control	2.58	3.52	4.71	0.11	0.20	3.96

Notes: Dependent variable is average of weekly visits (calculated by authors from gym data or stated by subjects in survey data) or exercise over the incentive period weeks one to four. There is one observation per subject for these regressions. Regressions include strata fixed effects. Robust standard errors are reported in parentheses. Cuts of the data above are based on our prespecified randomization stratification by target level of exercise by membership status.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Source: Authors' calculations

of nonmembers, 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.[31] The self-reported and computerized-record estimates of the incentive effects are similar in the subgroup panels B and C (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

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

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 overall effects in panel A, we see that this ratio is 82 percent for existing members and 63 percent for nonmembers (a weighted average effect of 70 percent). 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 nonmembers at/above their target exercise level, there appears to be considerable substitution.[32] Taken literally, 74 percent 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 percent), we did observe some statistically-significant differential attrition in survey response between the treatment and control groups among nonmembers as seen in online 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 nonresponse bias for nonmembers in such analyses, we estimate the degree to which nonresponse might affect our substitution estimates in online 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 percent. Overall since roughly 70 percent 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 appears to have 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).[33]

IV. Robustness and Heterogeneity

A. *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.[34] 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 the interdependence of exercise behavior among

[32] 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 that this group increased their overall exercise in response to the incentives.

[33] Of course, this conclusion assumes that company gym attendance is equivalent to exercise at the company gym (i.e., individuals are not going to the gym without exercising).

[34] 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.

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 preincentive 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 (2010) 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 preintervention to the intervention period that is related to the fraction of the control group's department that is incentivized. Moreover, only 3 percent of the company employees were incentivized at any one point in time.

B. *Heterogeneity of Response to Incentives*

In Table 6, 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 versus at/above target split, these cuts were not prespecified, so the results should be interpreted with caution. We motivate our heterogeneity analysis from the results on commitment take-up in Table 4, which suggested interesting patterns of commitment demand related to gender and weight. Surprisingly, we also found evidence that those with high levels of prestudy 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 6 are broadly in line with what one would expect given the analysis in Section IV. 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 the sexes there are much larger differences between the incentive-only and the incentive+commit groups in the posttreatment 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

Table 6—Heterogeneity Cuts on Treatment Effects

	In-treatment effect			Early posttreatment effect[a]		
Heterogeneity cut	Control mean[b] (1)	Treated pooled[c] (2)	Commitment take-up (3)	Incentive-only (4)	Incentive+commit (5)	*P*-value: (4) = (5) (6)
Panel A. Exercise						
Low preperiod exercise	0.03	0.17***	0.09	−0.00	0.07*	0.03
		(0.03)		(0.03)	(0.04)	
Middle preperiod exercise	0.11	0.23***	0.12	0.07**	0.13***	0.07
		(0.03)		(0.03)	(0.03)	
High preperiod exercise	0.41	0.11***	0.15	−0.02	0.03	0.17
		(0.03)		(0.04)	(0.04)	
Panel B. 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	0.39	0.13***	0.15	0.01	0.05	0.28
		(0.03)		(0.04)	(0.04)	
Panel C. 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)	
Panel D. Obesity						
Nonobese/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)	

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 nonmembers. The heterogeneity cut by prestudy exercise is based on terciles of self-reported days of typical exercise each week that subjects give in the initial survey.

[a] Early posttreatment period is weeks 5–13.

[b] Control mean reported for weeks 1–13.

[c] Indicates treatment with either incentive-only or incentive+commit treatment.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Source: Authors' calculations

incentive-only treatment. Taken together, these results provide at least suggestive evidence in support of our findings in Section IV that stronger effects of the incentive+commit treatment in the postperiod 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 prestudy exercise patterns. Here, in order to combine members and nonmembers effectively, we use answers to the initial survey question asking employees to report the typical number of days on which they exercise

per week and split the sample into terciles by that measure.[35] Column 3 confirms our discussion from Section IV that those who were exercising regularly prior to the study—as indicated by either exercise relative to target or prestudy exercise frequency—created commitments at rates similar to their lower exercise counterparts.[36] The posttreatment 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 postincentive period. This lack of change is not surprising, given their high rates of exercise prior to the study, but it does highlight the 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.

V. 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 show only very small lasting effects that dissipate 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 percent 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 percent 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 pretreatment exercise levels and our estimates of substitution effects, we conclude that approximately 35 percent of the cost of the incentive program was spent on new behavior, while 65 percent 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.[37]

[35] The cuts are 0–0.5 days, 1 to 3 days, and more than 3 days per week.

[36] For this table we split subjects based on tertiles of prestudy exercise using gym attendance for existing members of the gym and self-reported exercise frequency in the initial survey for nonmembers.

[37] 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.

More generally, the findings here suggest that programs incorporating temporary incentives with unincentivized periods 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 healthcare spending. There is reason to believe that behavioral changes of the magnitude induced by our program could pay for themselves through these mechanisms. For example, Jacobson and Aldana (2001) report that increasing exercise from zero to one or one to two days per week is associated with significant decreases in absences from work. 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. Our program cost $57 per employee. Based on that rate, an incentive program such as ours could pay for itself through reduced absences alone if roughly one in three employees experience one fewer absences per year.

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 and often higher than 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 a new commitment device. We believe this finding is an important insight for this literature as it suggests it is harder than one might have anticipated to ex ante identify individuals with the sort of time-inconsistency problems that lead to the demand for commitment. If people have access to alternative strategies for overcoming time inconsistency, then looking for troubling patterns of behavior or asking questions about self-control struggles (as we did in our surveys) may fail to identify people who have an underlying demand for commitment technologies. One particular 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; Baumeister, Muraven, and Tice 2000) and recent models of willpower depletion in economics (Ozdenoren, Salant, and Silverman 2012) suggest willpower 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)confident 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 overconfidence are ambiguous. We believe our findings here suggest that understanding the role overconfidence plays in the demand for commitment and the effects overconfidence 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. 2015. "Naiveté, Projection Bias, and Habit Formation in Gym Attendance." *Management Science* 61 (1): 146–60.

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): 635–72.

Augenblick, Ned, Muriel Niederle, and Charles Sprenger. 2013. "Working Over Time: Dynamic Inconsistency in Real Effort Tasks." National Bureau of Economic Research (NBER) Working Paper 18734.

Babcock, Philip, Kelly Bedard, Gary Charness, John Hartman, and Heather Royer. 2011. "Letting Down the Team? Evidence of Social Effects of Team Incentives." National Bureau of Economic Research (NBER) Working Paper 16687.

Babcock, Philip S., and John L. Hartman. 2010. "Networks and Workouts: Treatment Size and Status Specific Peer Effects in a Randomized Field Experiment." National Bureau of Economic Research (NBER) Working Paper 16581.

Baicker, Katherine, David Cutler, and Zirui Song. 2010. "Workplace Wellness Programs Can Generate Savings." *Health Affairs* 29 (2): 304–11.

Baumeister, Roy F., Ellen Bratslavsky, Mark Muraven, and Dianne M. Tice. 1998. "Ego Depletion: Is the Active Self a Limited Resource?" *Journal of Personality and Social Psychology* 74 (5): 1252–65.

Baumeister, Roy F., Mark Muraven, and Dianne M. Tice. 2000. "Ego Depletion: A Resource Model of Volition, Self-Regulation, and Controlled Processing." *Social Cognition* 18 (2): 130–50.

Benartzi, Shlomo, and Richard Thaler. 2004. "Save More Tomorrow: Using Behavioral Economics to Increase Employee Saving." *Journal of Political Economy* 112 (1): S164–87.

Beshears, John, James J. Choi, David Laibson, Brigitte C. Madrian, and Jung Sakong. 2011. "Self Control and Liquidity: How to Design a Commitment Contract." RAND Working Paper Series WR-895-SSA.

Brune, Lasse, Xavier Giné, Jessica Goldberg, and Dean Yang. 2013. "Commitments to Save: A Field Experiment in Rural Malawi." World Bank Policy Research Working Paper 5748.

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

Charness, Gary, and Uri Gneezy. 2009. "Incentives to Exercise." *Econometrics* 77 (3): 909–31.

Cox, D. R. 1958. *Planning of Experiments*. New York: John Wiley and Sons.

DellaVigna, Stefano, and Ulrike Malmendier. 2006. "Paying Not to Go to the Gym." *American Economic Review* 96 (3): 694–719.

Finkelstein, Eric A., Laura A. Linnan, Deborah F. Tate, and Ben E. 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): 981–89.

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): w822–31.

Fryer, Jr., Roland. 2011. "Financial Incentives and Student Achievement: Evidence from Randomized Trials." *Quarterly Journal of Economics* 126 (4): 1755–98.

Giné, Xavier, Jessica Goldberg, Dan Silverman, and Dean Yang. 2012. "Revising Commitments: Field Evidence on the Adjustment of Prior Choices." National Bureau of Economic Research (NBER) Working Paper 18065.

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.

Gneezy, Uri, Stephen Meier, and Pedro Rey-Biel. 2011. "When and Why Incentives (Don't) Work to Modify Behavior." *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." National Bureau of Economic Research (NBER) Working Paper 16624.

Jacobson, Bert H., and Steven G. Aldana. 2001. "Relationship Between Frequency of Aerobic Activity and Illness-Related Absenteeism in a Large Employee Sample." *Journal of Occupational and Environmental Medicine* 43 (12): 1019–25.

Jeffery, Robert W., Wendy L. Hellerstedt, and Thomas L. Schmid. 1990. "Correspondence programs for smoking cessation and weight control: A comparison of two strategies in the Minnesota Heart Health Program." *Health Psychology* 9 (5): 585–98.

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

Just, David R., and Joseph Price. 2013. "Using Incentives to Encourage Healthy Eating in Children." *Journal of Human Resources* 48 (4): 855–72.

Karlan, Dean, and Leigh L. Linden. 2014. "Loose Knots: Strong versus Weak Commitments to Save for Education in Uganda." National Bureau of Economic Research (NBER) Working Paper 19863.

Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan. Forthcoming. "Self-Control at Work." *Journal of Political Economy.*

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics* 112 (2): 443–78.

List, John A. 2009. "An introduction to field experiments in economics." *Journal of Economic Behavior and Organization* 70 (3): 439–42.

Loewenstein, George, Ted O'Donoghue, and Matthew Rabin. 2003. "Projection Bias in Predicting Future Utility." *Quarterly Journal of Economics* 118 (4): 1209–48.

Milkman, Katherine L., Julia A. Minson, and Kevin G. M. Volpp. 2014. "Holding the Hunger Games Hostage at the Gym: An Evaluation of Temptation Bundling." *Management Science* 60 (2): 283–99.

O'Donoghue, Ted, and Matthew Rabin. 1999. "Doing It Now or Later." *American Economic Review* 89 (1): 103–24.

O'Donoghue, Ted, and Matthew Rabin. 2001. "Choice and Procrastination." *Quarterly Journal of Economics* 116 (1): 121–60.

Ozdenoren, Emre, Stephen W. Salant, and Daniel Silverman. 2012. "Willpower and the Optimal Control of Visceral Urges." *Journal of the European Economic Association* 10 (2): 342–68.

Royer, Heather, Mark Stehr, and Justin Sydnor. 2015. "Incentives, Commitments, and Habit Formation in Exercise: Evidence from a Field Experiment with Workers at a Fortune-500 Company: Dataset." *American Economic Journal: Applied Economics.* http://dx.doi.org/10.1257/app.20130327.

Strotz, R. H. 1955–56. "Myopia and Inconsistency in Dynamic Utility Maximization." *Review of Economic Studies* 23 (3): 165–80.

Volpp, Kevin G., Leslie K. John, Andrea B. 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): 2631–37.

Volpp, Kevin G., Mark V. Pauly, George Loewenstein, and David Bangsberg. 2009a. "An Agenda for Research on Pay-For-Performance For Patients." *Health Affairs* 28 (1): 206–14.

Volpp, Kevin G., Andrea B. Troxel, Mark V. Pauly, Henry A. Glick, Andrea Puig, David A. Asch, Robert Galvin, et al. 2009b. "A Randomized, Controlled Trial of Financial Incentives for Smoking Cessation." *New England Journal of Medicine* 360 (7): 699–709.