NBER WORKING PAPER SERIES

CAN FINANCIAL INCENTIVES HELP PEOPLE TRYING TO ESTABLISH NEW HABITS? EXPERIMENTAL EVIDENCE WITH NEW GYM MEMBERS

Mariana Carrera
Heather Royer
Mark Stehr
Justin Sydnor

Working Paper 23567
http://www.nber.org/papers/w23567

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2017

We are thankful for funding from the UPenn Roybal Center and a UCSB Faculty Senate Grant. We appreciate the outstanding research assistant work of Chang Lee, Garrison Schlauch, Paul Fisher, Jordan Hsieh, Alan Thomas, Rachael Collins, Abigal Whited, Fred Li, Anjuri Kakkar, Pooja Padmakumar, Madeline Thomas, Precious Adeleye, Jaelynn Theobalds, Emma Chelala, and Angeline Xiong. We are thankful for the comments and suggestions of Julien Mousqués along with those of various seminar and conference participants. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Can Financial Incentives Help People Trying to Establish New Habits? Experimental Evidence with New Gym Members

Mariana Carrera, Heather Royer, Mark Stehr, and Justin Sydnor

NBER Working Paper No. 23567
July 2017
JEL No. C93,D03,I12

ABSTRACT

We conducted a randomized controlled trial testing the effect of modest incentives to attend the gym among new members of a fitness facility, a population that is already engaged in trying to change a health behavior. Our experiment randomized 836 new members of a private gym into a control group, receiving a $30 payment unconditionally, or one of 3 incentive groups, receiving a payment if they attended the gym at least 9 times over their first 6 weeks as members. The incentives were a $30 payment, a $60 payment, and an item costing $30 that leveraged the endowment effect. These incentives had only moderate impacts on attendance during members' first 6 weeks and no effect on their subsequent visit trajectories. We document substantial overconfidence among new members about their likely visit rates and discuss how overconfidence may undermine the effectiveness of a modest incentive program.

Mariana Carrera
Case Western Reserve University
Weatherhead School of Management
11119 Bellflower Rd. #274
Cleveland, OH 44106
mpc67@case.edu

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

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

Justin Sydnor
Wisconsin School of Business
ASRMI Department
University of Wisconsin at Madison
975 University Avenue, Room 5287
Madison, WI 53726
and NBER
jsydnor@bus.wisc.edu

A online appendix is available at http://www.nber.org/data-appendix/w23567

Exercise is a prototypical example of a positive health behavior where self-control problems appear to lead to suboptimal establishment of habits. Only 21% of Americans get the recommended amount of weekly exercise[1] and many people feel they exercise less than would be optimal (e.g., Royer, Stehr and Sydnor, 2015). Many also pay substantial fees over long periods for gym memberships that they do not use (DellaVigna and Malmendier, 2006).

A small recent literature has used randomized-controlled trials to test whether temporary financial incentives for exercising can help people establish more lasting habits. While the literature has identified that temporary incentives for going to the gym have impacts in some cases (especially Charness and Gneezy, 2009), the effects tend to be modest and do not always appear (Charness and Gneezy, 2009; Acland and Levy 2015; Royer, Stehr and Sydnor, 2015; Carrera et al., 2017; Rohde and Verbeke, 2017).

Financial incentives may be less effective when they target a broad population whose potential for habit formation is not at its peak. Behavior change that leads to successful habit formation may require that people are motivated and prepared for making a change (Ajzen, 1985). Recent research finds, for example, that many people start new exercise routines and make other life changes on salient dates, such as birthdays (Dai, Milkman, and Riis, 2014). That pattern suggests that it may be advantageous to attempt behavior change on specific dates when motivation might be high. Perhaps, then, timing incentive programs to coincide with moments when people have already taken the first step toward establishing a new habit could be more effective. On the other hand, however, people who are attempting to establish new habits may generally be overoptimistic and facing other adjustment challenges that may overwhelm any effects of extrinsic incentives.

We test the effects of financial incentives that coincide with endogenous attempts at establishing new habits using a randomized controlled trial with new members of a gym. This is a useful group to study in this context because they have all already engaged in costly actions – paying membership fees and going through the enrollment process – that signal an intention to use the gym. Prior research also shows clearly that many people who join gyms fail to establish a gym-going habit. These patterns of attendance are clear amongst our study population. New

[1] https://www.cdc.gov/physicalactivity/data/facts.htm

members of the gym report that they plan to attend the gym 3 times per week. In reality, in absence of an intervention, visits initially start at 2 visits per week in the first week and fall quickly to an average of only 1 visit per week by the end of the second month of membership. Thus, both the fact that new members have shown that they are ripe for pursuing behavioral change and the fact that they often face difficulty in establishing and maintaining their exercise habits make this population an interesting and potentially fruitful one to study.

Our experiment randomized all 836 new members who enrolled in a gym over the course of an 8 month period into one of four arms: a control group and 3 incentive groups. For all of the incentive groups, subjects earned incentives by attending the gym at least 9 times in the first 6 weeks of membership (i.e., an average of 1.5 visits per week). We chose 9 visits as the target because it was the median number of visits for members who joined prior to our intervention. The median seemed appropriate because it was high enough to be potentially motivating for many members yet not so high as to be unrealistic. Among the incentive groups, subjects in one group earned $30 in an Amazon.com gift card for reaching the threshold, subjects in one group similarly earned $60 in an Amazon.com gift card, and subjects in the last group earned a specific but subject-chosen item sold by Amazon.com worth approximately $30. Research on the endowment effect (Kahneman, Knetsch and Thaler, 1990) inspired this item-based incentive. We hypothesized that selecting a specific item at the outset might create a sense of ownership for that item so that not hitting the target visit rate might feel like a "loss" of the item. For loss-averse people, that could make an incentive based on an item more powerful than an equivalent-valued monetary prize, even though the monetary prize is more fungible.

Our experiment reveals that additional incentives for visits early during a new gym membership were only moderately effective at helping people to increase their exercise. Across all of the incentive treatments we find only small effects on the number of visits over the first 6 weeks of membership and no effect of having an incentive on visit rates after that incentive period. We find that the item incentive induces slightly more visits than the equivalent-valued monetary incentive, but the differences are modest and not statistically significant. In heterogeneity analysis, not surprisingly, low exercisers and those who struggle with establishing

exercise habits had the largest increases in total visits in response to the incentives. Overall we conclude that the provision of modest additional financial incentives only marginally changed the behavior of new gym members.

2. Experimental Design

2.1 Setting

Our experiment took place at a commercial gym consisting of roughly 3,000 members in a large Midwestern city between September 2015 and April 2016. The gym is affiliated with a local nearby university but is open to the public and is separate from the campus' primary student fitness facility. In our study sample, 49 percent are associated with the University in some way as faculty, staff, or students. The baseline membership cost is $59 per month. However, membership discounts are available to a number of groups including those associated with the university.

The operation of the facility is rather typical for a fitness gym. The gym is open 7 days per week from 5:30 a.m. to 12:30 a.m. on weekdays and 8 a.m. to 10 p.m. on the weekend. Members have an ID card that is swiped by front desk personnel upon entry to the facility. These timestamped entry records form the primary data for this study.[2]

We do not track the specific activities people engaged in while at the gym, only the number of days they came to the gym and checked in. One potential concern with login records as the outcome measure for an incentive is that it may encourage people to show up at the gym only to "swipe in" but not to exercise in their normal way. In general, we were not overly worried about inducing this type of behavior because there are real costs (e.g., time, parking) associated with accessing the gym for most people, which tend to reduce the likelihood of this type of behavior. We also introduced a new checkout procedure partway through the study (in February 2016). Participants after that time were required to swipe out after attending the gym for at least 10 minutes in order to get credit for a visit toward their incentive. Introducing this procedure did not change visit patterns or the estimated treatment effects in the study and the

[2] In the event that a member forgets her ID card, the staff will look her up in the computer to log the entry, which still appears in the same timestamped records.

swipe-out records reveal that the vast majority of gym visits lasted substantially longer than 10 minutes.

2.2 Recruitment and treatment assignment

Our subject pool is new members. Upon enrolling with the gym, members fill out a membership packet. We embedded our experimental randomization into this enrollment process by attaching our study enrollment forms at the end of each new membership packet during the study period (see Appendix for a copy of study enrollment forms). The enrollment forms began with a flyer highlighting their randomized treatment assignment.[3] We used a simple randomization procedure where enrollment forms were sorted in a stack that alternated control and each of the three treatment assignments. Gym staff simply used the enrollment packet on the top of the stack as new members joined the gym.[4]

The enrollment forms included an IRB-approved consent form, short survey, and contact information sheet. Subjects had to consent to participate in the study in order to receive payment. From the membership packets, there was a record of the treatment offered and whether or not the new member chose to participate in our study. For those who consented to participate, we can match at the individual-level their survey data from the enrollment packet and their gym attendance record. In principle, if new members did not selectively choose whether to participate, our analysis of consenters (or what we later refer to as participants) is sufficient and will lead to unbiased estimates of the treatment effects on the treated. However, we want to test whether this assumption of non-selectivity is true. From the gym, we obtain two useful data: a) the fraction of consenters across the four arms of the study and b) the average rate of attendance for each of the four arms unconditional on consenting to participate (since the gym keeps visit data for all members).

We randomized members into one of four groups. These groups were as follows:

[3] For example, for the $30 incentive group, the flyer included "You will be eligible for a $30 Amazon.com gift card" and "You get the gift card as long as you visit [the gym] on at least 9 days over your first 6 weeks as a member."

[4] There are three membership types at the gym – regular, graduate student, and those who signed up through a well-being improvement company affiliated with their health insurance plan. We randomized treatment assignments within stacks of membership forms for these three groups separately. As such, our randomization is stratified by membership type.

(a) Control group: received a $30 Amazon gift card after six weeks unconditionally.

(b) Money $30 group: received a $30 Amazon gift card if attended gym at least 9 days in their first 6 weeks of membership.

(c) Money $60 group: received $60 Amazon gift card if attended gym at least 9 days in their first 6 weeks of membership.

(d) Item group: received a self-chosen item worth approximately $30 from Amazon if attended gym at least 9 days in their first 6 weeks of membership.

For Money $30, Money $60, and Item groups, the receipt of their prize was conditional on their attendance. The control group received a $30 payment simply for participation (e.g., enrollment survey completion), which ensured that any observed effects of the incentive were not caused by differential "good-will" effects between the treatment groups and control. Providing a payment to the control group also ensured similar participation rates in the full study (e.g., consenting to complete the initial survey).[5]

Those randomized into the Item group were given the choice of one item among ten pre-selected products sold on Amazon for roughly $30. At the time of enrollment, we presented participants with the details of each of the products, including pictures and ratings of them. At that time, subjects were asked to select one of the products as a prize and were told (truthfully) that the product would be ordered and held for them until they completed their 6th week of membership. Of course, the actual receipt of the prize was conditional on attending the gym at least 9 days over those 6 weeks. One of the item sheets is displayed in the Appendix.[6]

Our selection process for these 10 products started with collecting a long list of products available at Amazon.com for prices ranging $27-35 since prices fluctuate frequently on Amazon, with at least 100 reviews and an average star rating of at least 4. We then did extensive polling

[5] We cannot rule out the possibility that the unconditional payment of a $30 gift card affected the behavior of our control group and that our incentives would have appeared to have larger effects when compared to a control group that did not receive any payment. Prior studies, however, have not found that the effects of a gym incentive depend on whether the control group is uncompensated or receives an unconditional reward (Charness and Gneezy, 2009, Rohde and Verbeke, 2017)

[6] Item group subjects had a choice over the following products: Ninja Master Prep blender (36%), Play X Earbuds (25%), Bluetooth Shower Speaker (9%), Portable 8W Solar Charger (8.4%), a portable hammock (7.8%), an electric kettle (6.6%), wireless desktop keyboard/mouse combination (3%), Google Chromecast (1.8%), Redragon gaming mouse (1.8%) where the numbers in parentheses represent the frequency with which those items were chosen.

on Mechanical Turk to choose the items that generated the most interest as potential prizes for a study. In the end, the average price paid per item was $31.53.

Classical economic theory predicts that this type of item incentive would be perceived as (weakly) less valuable than an unconstrained monetary prize of the same value. The motivation for this incentive design is the endowment effect (Kahneman, Knetsch and Thaler, 1990), a phenomenon where people appear to value objects much more if they feel a sense of ownership for them. The idea in our context was that if individuals felt strong attachment to their chosen item, this incentive may work better than a monetary incentive with the same value. We tried to instill a sense of ownership at the beginning of the experiment by emailing subjects a picture of their item twice: once to confirm their choice of item and its order, and again after it had arrived, using a Post-It note to label it with their name and telling them that their item was waiting for them at the gym (all of which was true). Contrasting the Item group with the Money $30 group and Money $60 group allows us to place a monetary value on the endowment effect.

2.3 Summary statistics

Table 1 presents descriptive statistics for the full sample of new members who joined the gym during our study period. The first two columns show the overall mean and the control group means for several key variables – first, the participation rate (the fraction of individuals consenting to be a part of our study) and second, variables collected by the gym for all members. The next three columns show the difference between the control group and each of the three treatment groups. The final column presents p-values testing whether each of the treatment assignments have equal means.[7]

Nearly all new members consented and were eligible to participate in the study (over 80%). There are slight differences in participation rates across the experimental arms but we are unable to reject that the participation probability is the same across the treatments. Approximately half of the new members are associated with the nearby university and nearly

[7] These p-values come from tests of equivalence of the treatment dummy coefficients in OLS regressions of each of the variables listed in the far right column of Table 1 on the treatment status indicators. All models are estimated with heteroscedasticity-consistent standard errors.

all of those who are university affiliated are students. Our randomization appears to be balanced as none of the p-values in the final column indicate statistically significant differences by treatment status.

Although the packets were distributed in equal numbers, there was some random variation in final sample sizes for the treatment assignments caused by some packets being given to potential members who never joined, or to ineligible members (e.g., existing members completing enrollment forms to change their membership type).

Table 2 parallels Table 1 but lists characteristics for those who consented to participate in the study, and for whom we have survey measures from the new membership packets. Across all measures, at standard significance levels, we cannot reject that the means are the same across treatment groups, suggesting that the treatment groups were balanced. The average age of participants in the study was 35, similar to the overall sample of new members, and similar across treatment groups. Compared to the full sample of new members, participants were slightly more likely to be female (58% vs 55%). Among the participants, we observe that those in the item group were 7 percentage points less likely than control to be female, a slight gender imbalance. Overall nearly 90% of participants report having a college degree or an advanced graduate degree. The high education rates of our sample are consistent with the gym's target population of university staff, faculty, graduate students and hospital employees.

The new membership packet survey also asked participants to report some basic information about their past exercise behavior and their expectations for visit patterns at the gym. When respondents answered the survey they were already aware of their treatment assignment, and as such the answers to these questions could be influenced by treatment expectations. Overall 43% of participants reported that they had exercised on average one day or less per week over the prior year. The frequency of reporting low prior-year exercise was higher for participants in the incentivized treatment groups relative to the control group. We also asked participants to characterize their past experience establishing an exercise routine. Subjects could choose from 5 statements the one that best characterized their experience and

three of these indicated a failure to establish an ongoing exercise habit in the past.[8] The majority (55%) of subjects chose one of these three options. Consistent with the patterns for exercise frequency, those in the treatment groups were a little more likely than control to state they did not have a prior exercise routine.

The new members expected to attend regularly. On average, new members predicted that they would visit the gym 3 times per week. Interestingly, that expectation was not significantly different for those in the incentive groups even though they were aware of the incentive opportunity. New members were also overall quite confident that they would visit the gym at least 9 times over their first 6 weeks as members. Participants assessed their likelihood of attending at least 9 times and using their responses we estimate that overall members believed they had around a 78% chance of meeting the 9 visit-per-week target.[9] Again, interestingly, this was similar for those in the treatment and control groups.

3. Results

3.1 Results for participants

We begin our analysis by examining the patterns of visit rates over the first 14 weeks of membership for the 690 new members who participated in the study. Figure 1a shows these patterns for the control group and the three incentive groups pooled together.

The most glaring pattern is the downward trend in visits over time for new members. This highlights why an intervention aiming to encourage members to come frequently or consistently might be useful. Eight weeks into their membership, new members go half as frequently as they did in the first week of their membership. The dashed line reveals that the control group visit rates fell from an average of 2 visits during the first week of membership to around 1 visit per week by the end of the second month of membership. Recall that members

[8] The specific options were: "I have never tried to establish an exercise routine" (10%); "I have repeatedly tried to establish an exercise routine, but have never been successful" (10%); "I have at times established a regular exercise routine, but have been unable to stick to it for long periods" (35%); "Although I struggle with my commitment occasionally, I am usually able to keep up a regular exercise routine" (34%) and "I am a workout buff: I keep a regular exercise routine without much problem at all" (11%).

[9] The survey question appears in the appendix. We assigned 10%, 30%, 50%, 70% and 90% to those responding 0-20%, 21-40%, 41-60%, 61-80% and 81-100% respectively.

expected themselves to make about 3 visits per week, suggesting that without added incentives, they attended about one third as often as they desired by two months after joining.

The average visit rates for the members assigned to one of the three incentive groups were somewhat higher over the first 6 weeks of membership, consistent with the presence of incentives. The differences, however, were quite modest. On average, across the first six weeks, the members assigned to the incentive treatments made 0.14 more visits per week than the control group. The largest difference was in week 2 when the control group made an average of 1.5 visits while the incentivized groups averaged 1.73 visits. Interestingly, there was a modest bump in visit rates for the incentivized group in the 6^{th} week of membership, which was the last week during which this group could make visits to count toward the 9 visit incentive threshold. From week 10 on, the average visit rates for the incentivized groups and control groups were very similar and hover around 1 visit per week.

Figure 1b shows these patterns separately for each of the three incentive groups. The patterns look generally similar. In all cases we see the same sort of sharp decline in attendance rates over the first two months of membership. The average visit rates for the Item group were higher than those of control in all but week 10 suggesting that the Item incentive might have had a larger effect on attendance than the other incentives. However, caution is warranted in interpreting these raw visit differentials, as the modest selection patterns identified in Tables 1 and 2 could be partly responsible for these results. Below we present regression results to quantify the difference in visit rates and to control for the modest differences in observables between subjects in the different treatments.

It is not obvious that the average number of visits measure examined in Figures 1a and 1b is the most relevant since the incentives were threshold-based and the number of visits distribution has a non-negligible and long right tail. Thus, it is useful to analyze the distribution of the number of visits made over the first 6 weeks. Figure 2 shows histograms with the number of visits top-coded at 24 (average 4 per week) due to the long and sparse right tail in visit counts. The dashed line in each graph denotes the 9-visit incentive target.

The histograms reveal a few interesting patterns. For all groups, there is considerable diversity in the number of visits new members make during the first 6 weeks, but in general

most of the mass lies below 12 visits (i.e., 2 visits per week on average). For the Control group the highest peaks in the histogram occur between 2 and 7 visits, and the overall average is 9.41. The Item incentive group's visits were shifted slightly to the right with an overall average of 10.75. A two-sample Wilcoxon rank-sum test of the difference of the two distributions has a p-value of 0.09.

Both of the monetary incentives, but especially the Money60 treatment, show some evidence of "hollowing out," with both more mass at visit rates above 9 and below 3 than the control. For the Money60 incentive there is a distinct peak in the histogram at 10 visits in the first 6 weeks and hollowing out of the mass around 7 visits relative to what we see for the control group. However, the average visits for the Money30 and Money60 treatments were only modestly higher than the Control average at 9.99 and 10.09, respectively. This is because the increased fraction attending 9 or 10 times was offset by a higher fraction of members in the monetary groups who visited only once during the incentive period. One possible interpretation of this "hollowing out" pattern for the Money60 incentive is that the higher monetary treatment might have led to some discouragement among a subset of new members and caused them to give up attending earlier. We caution, however, that overall we cannot detect statistically significant differences in these distributions: the p-values on the Wilcoxon rank-sum test of the difference of distributions between Control vs. Money30 and Control vs. Money60 are 0.83 and 0.41, respectively.

In Table 3 we present regression results to quantify the average treatment effects observed in Figures 1 and 2. For these regressions we run models of the form:

$$y_i = \alpha + \beta D_i^{treatment} + X_i'\theta + \varepsilon_i,$$

where y_i is a measure of visits and $D_i^{treatment}$ is an indicator that takes the value of 1 for individuals in the treatment group and 0 otherwise. In Panel A, we present regressions pooling all three incentive treatments together to estimate a single treatment effect. In Panel B, we estimate three separate treatment coefficients, one for each of the treatment groups relative to control. The three visit measures used as dependent variables are a dummy variable for meeting the 9-visit threshold over the first 6 weeks, the number of visits in the first 6 weeks,

and the number of visits in weeks 7-12 (a test of the lasting effects of the intervention, an interest in prior literature (Charness and Gneezy, 2009 and Royer, Stehr, and Sydnor, 2015)).

We consider models with and without controls. In principle, such controls are not necessary due to the randomization of the treatments, but in some cases, there are slight differences in these covariate means across groups, so for robustness, we also include these control variables. Qualitatively, the addition of the control variables has little impact on treatment effect point estimates. The matrix of controls, which we include in even-numbered columns in both tables, includes age in years, an indicator for being female, having a university affiliation, dummies for special membership type (e.g., student, senior), and indicators for self-reported frequency of exercise in the year prior to joining the gym and for reporting no success establishing an exercise routine in the past. Throughout we run ordinary least squares regressions with heteroscedasticity-consistent standard errors.

The coefficient estimates in columns 1 and 2 of Table 3, Panel A, show that members facing an incentive were 9-10 percentage points more likely to reach the threshold of nine visits in the first six weeks than control group subjects. This result is statistically significant and substantial relative to the control group's 48 percent probability of meeting the threshold. The increase in the average number of visits, however, is less pronounced and only marginally statistically significant with controls. In column 3, without controls, the estimated increase of 0.85 visits over the first six weeks is equivalent to the sum of the differences between the dashed and dotted lines, over weeks 1 to 6, in Figure 1a. When controls are added, the estimated treatment effect increases slightly, to 0.98, and is significant at the 10% level. Columns 5 and 6 show that in the post-incentive period, weeks 7-12, the estimated difference in visits between the pooled incentive and control groups is smaller and not statistically significant, in line with the convergence of the dashed and solid lines seen in Figure 1a.[10]

Panel B presents the same regression estimates for each treatment group separately. Of the three incentive groups, Money 60 had the largest and most significant increase in the

[10] Our study was not powered to detect small post-intervention treatment effects. Our power calculations, based on the visit data of new members prior to our study, implied that with at least 150 participants in each group, we would have power to detect differences of 1.72 visits over the 6 week intervention period between two groups, or a 0.29 difference in average visits per week. Note that this minimum detectable difference is less than half as large as the effect on average weekly visits estimated by Charness and Gneezy (2009), for a threshold incentive of $100.

probability of meeting the 9 visit threshold, a 12 percentage point increase. This is not surprising since members of this group had the strongest incentive to meet the threshold. It is more surprising, however, that Money 60 also had the smallest increase in average visit rates after controlling for covariates. This reflects the apparent "discouragement effects" seen in Figure 2. Compared to the distribution of visits in the control group, the $60 incentive has more mass just above the threshold of nine visits, but also more mass far below the threshold, with the latter potentially representing people who visited *less* than they would have in the absence of the incentive. Thus, the average visit rate is only slightly larger in Money 60 versus the control group, despite a substantial increase in the probability of being above the threshold.

The estimated treatment effects of Item and Money30 are positive but statistically insignificant. Examining the point estimates across the different specifications, it is not immediately clear whether the Item or Money30 treatment is more effective. Recall that our goal in including these two treatments was to test for the endowment effect. Classical economic theory would predict that a fungible $30 is a weakly stronger incentive than a fixed item worth $30, but if the anticipation of owning a chosen item evokes "the endowment effect," then the Item treatment may have a stronger effect. All of these coefficients, however, are statistically insignificant at the 5% level, and only the effect of Item on 9+ visits is significant at the 10% level. The magnitude of Item treatment effect on 9+ visits is not negligible - nearly a 20% increase in the probability of attending 9 or more days.

3.2 Robustness check using assignment to treatment offer

Since our study enrollment packets contained information about incentives, individuals in the recruitment pool could learn about their assigned treatments before deciding whether to participate. In Table 1, we showed that we cannot reject the null hypothesis that treatment status had no effect on participation rates. Nonetheless, in this section we address the possible concern of differential selection by treatment group by conducting a simple intent-to-treat analysis.

While we do not have survey data for those who did not participate, our agreement with the gym does allow us to calculate visit rates for the full sample of new members. We can

compute visit outcomes for all members invited to participate and test for mean differences between the groups offered different treatments.

Table 4 summarizes visit outcomes for all who were invited to participate in the study, by their assigned treatment group or "treatment offer." The differences reported between each group and the control group are "intent to treat" effects. These synthesize the same information we present in Table 3 except we present means and differences in means. The second panel shows the means of each visit outcome when all treatments are pooled into a single incentive group. The means for 9+ visits and visits over 1st 6 weeks are larger than the means of the control group, showing a marginally significant 0.07 percent increase in meeting the 9-visit threshold and an insignificant increase of 0.57 in average visits among new members who were invited to join an incentive group relative to those invited to the control group. It is not surprising that these impacts are smaller than the analogous estimates in Table 3 because the treatments should have little effect on the non-participants – leading to a dampened effect overall when we combine participants and non-participants. The remaining panels of Table 4 show means by specific treatment group. The differences among the treatment groups follow the same patterns seen in columns 1, 3, and 5 of Table 3, Panel B and discussed in the previous section. This analysis is less powerful given the non-participant rates but is consistent with our earlier analysis and further indicates that the main results are not driven by selection.

3.3 Heterogeneity

The overall effects presented thus far may mask interesting and substantial heterogeneity – especially given the diversity in the new member population. In this light, we investigate whether the treatment effects among participants differ by survey measures of exercise frequency and familiarity with maintaining an exercise routine. A threshold incentive might work better for those with low levels of exercise and little experience maintaining an exercise routine since these groups presumably have more scope for improvement in their exercise habits. Alternatively, if the goal is not realistic, then a threshold incentive may work better for those with higher levels of exercise and more experience maintaining a routine,

particularly if their exercise level in the absence of the incentive was close to but did not exceed the threshold.

Table 5 presents the results of heterogeneity cuts along these lines. We define as low exercisers those who reported exercising one or fewer times per week in the year before joining the gym and high exercisers as those who reported exercising two or more times per week. We categorize individuals as unsuccessful in maintaining an exercise routine if they report on our survey that they have never tried to establish an exercise routine, have been unsuccessful in trying to establish an exercise routine, or have been unable to sustain an exercise routine for a long period of time. We categorize individuals as successful if they are usually able to maintain an exercise routine or do so without much trouble.

Overall we do not detect substantial heterogeneity in the treatment effects by these cuts. However, we are limited in power to detect differences across the groups. The incentive effect point estimates on encouraging people to exceed the 9-visit threshold are elevated for those with more successful previous exercise routines, which may reflect the fact that this group was more likely to be near that threshold to begin with. The effects on average visit rates during and after the intervention, though, are higher for low exercisers and those with less prior success with exercise.

4. Discussion and Conclusion

We conclude that the provision of moderately-sized financial incentives only moderately helped new gym members establish better habits for using the gym. This suggests that, at least for exercise, timing financial incentives to align with endogenous attempts at behavior change may not be the most successful strategy for improving exercise habits.

One question raised by our results is whether the small effects of our incentives are related to their size and threshold nature. The results here do not rule out that a different type of incentive structure, like a per-visit incentive, or higher incentive stakes, might have generated stronger behavior change. However, two prior gym incentive studies, Charness and Gneezy (2009) and Acland and Levy (2015), documented substantial average response to threshold-based incentives and even saw some lasting effects on attendance once the

incentives were removed.[11] So, ex ante, there was no reason to expect that a threshold incentive would be ineffective amongst our study population. The incentive stakes we used in this experiment, though, are smaller than those in previous studies, with our $60 treatment offering stakes around half the size of those in Charness and Gneezy (2009).[12] Also, our "monetary" treatments offered Amazon gift cards rather than cash, which might make them slightly less valuable to some participants.[13] It is possible that doubling the size of our larger incentive might have had a more substantial impact on behavior. However, linearly extrapolating from the comparison of our $30 and $60 treatments, we might expect some increase in the probability of hitting the 9-visit threshold (a 20 percentage point treatment effect as compared to the 12 percentage point effect seen earlier) if we doubled the size of the largest incentive to $120, but no increase in the impact on overall visits, since the point estimate for the $60 treatment on visits is smaller than that for the $30 treatment. What we can say clearly is the promise of moderately-sized incentives in helping people establish exercise habits is limited. Modest-sized incentives (as opposed to high-powered incentives) are relevant when considering broad interventions, and are of the size firms and gyms often use.

Our results offer some interesting insights on our population of new members. New members are extremely overoptimistic about how often they will visit the gym, and there is a fast decline in their visit frequency over the first few months of membership. In our survey, 95 percent of participants indicated that they expected to visit the gym more than once per week on average, but the share of participants who did so was 63 percent in the first month and dropped to 34 percent in the third month. We also observe substantial dispersion in early-membership visit patterns. While some in the control group reach the 9 visit threshold as early as their second week, 14 percent don't come at all in weeks 2-4, and 28 percent make less than

[11] Among studies focusing on incentives for health behaviors and outcomes, some use threshold-based incentives (e.g., Acland and Levy, 2015; Babcock and Hartman, 2010; Babcock et al., 2015; Charness and Gneezy, 2009; John et al., 2011;, Rohde and Verbeke, 2017; Volpp et al., 2008; and Volpp et al., 2009) and some use more continuous-based incentives (e.g., Carrera et al., 2017, Cawley and Price, 2013; and Royer, Stehr, Sydnor, 2015).

[12] Charness and Gneezy's Study 1 offered participants $100 for attending the gym 8 times over 4 weeks compared to our incentive treatment of $60 for 9 visits over 6 weeks.

[13] The survey included a question "How often do you shop on Amazon.com?" The majority, 54% of respondents, chose "Frequently," 39% chose "Occasionally," and only 6.7% chose "Never or very rarely." Also, even our smaller gift card, $30, was enough to meet Amazon's minimum spending to obtain free shipping. Thus, we are not too concerned that participants would value the gift cards at less than their nominal value.

one visit per week, on average, in the first month.[14] Our incentives changes this distribution of visits only in a very localized way, inducing those just below the threshold to make a few extra visits over the six weeks but not raising average visit rates significantly.

One of the takeaways from this study is that future interventions aimed at closing the gap between intended and actual behavior among new members may need to be based on a better understanding of both their overconfidence and the rapid decline of their visit rate. For example, those who attended very infrequently over their first months of membership believed at the start that they were very likely to visit often and did not believe that their likelihood of visiting often would be influenced by the incentive.[15] That pattern is consistent with the possibility that overoptimistic people derive little additional motivation during the beginning of incentive programs because they (wrongly) believe they are very likely to earn the incentives.

Ultimately, we believe these results suggest that simply timing incentives to coincide with intrinsic motivation for change is likely to be insufficient on its own to help people reach their health goals. Even amongst new members, there is substantial heterogeneity in their past exercise habits. Instead of focusing on this group as a whole, it may be better to find ways to tailor incentives so that they are providing motivation on the margin for each individual. Moreover, tackling individuals' overconfidence by helping individuals set realistic and reasonable goals for themselves may make incentive programs more effective. In general, tailoring incentives is challenging, but in a population that has just started trying to change their own behavior, it may be more fruitful to add an extrinsic incentive after a short delay, or to design an adaptive incentive that adjusts based on the patterns of early success or failure observed. We see these as promising avenues for future research.

[14] This dispersion, which we anticipated when designing this experiment, highlights one reason why we chose a threshold as opposed to a per-visit incentive: Per-visit incentives result in the bulk of the budget being used to reward people who would be frequently visiting the gym anyway.

[15] Their perceived probabilities of making 9+ visits, reported in the initial survey after learning their treatment assignments, did not differ significantly between the treatment and control groups. (Average perceived probabilities imputed from a five-point scale ranged from 77.7% in the control group to 79.9% in the item group).

References

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

Ajzen, Icek. 1985. "From Intentions to Actions: A Theory of Planned Behavior." *Action Control*. Springer Berlin Heidelberg, 11-39.

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

Carrera, Mariana, Heather Royer, Mark Stehr, and Justin Sydnor. 2017. “The Structure of Health Incentives: Evidence from a Field Experiment.” *NBER Working Paper* No. 23188.

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.” *Econometrica,* 77(3): 909–931.

Dai, Hengchen, Katherine L. Milkman, and Jason Riis. 2014. "The Fresh Start Effect: Temporal Landmarks Motivate Aspirational Behavior." *Management Science* 60(10): 2563-2582.

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

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): 621-626.

Kahneman, Daniel, Jack L. Knetsch, & Richard H. Thaler. 1990. “Experimental Tests of the Endowment Effect and the Coase theorem.” *Journal of Political Economy*, 98(6): 1325-1348.

Rohde, Kirsten I.M., and Willem Verbeke. 2017. “We Like to See You in the Gym—A Field Experiment on Financial Incentives for Short and Long Term Gym Attendance.” *Journal of Economic Behavior & Organization*, 134: 388-407.

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

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." *The Journal of the American Medical Association* 300(22): 2631-2637.

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

Table 1. Participation Rates and Demographics for Full Sample of New Members

	Overall Mean	Control Mean	Item Difference	Money30 Difference	Money60 Difference	P-value of All Treatments=0
Participation Rate	0.83	0.85	-0.07	-0.03	0.00	0.22
Age	35.3	35.1	-0.05	0.00	0.62	0.96
	[14.6]	[14.5]				
Female	0.55	0.55	-0.05	0.04	0.04	0.28
University Affiliated	0.47	0.47	0.00	-0.01	-0.02	0.98
Student	0.44	0.44	0.02	0.00	-0.04	0.65
Secondary on Account	0.07	0.08	-0.03	-0.01	0.00	0.68
Number of Observations	836	207	200	215	214	

Notes: The overall mean column is the mean for the entire sample. The control mean column is the mean of the control group. The next three columns show the mean difference for the variable between the respective incentive groups and the control group. The p-value column displays the p-values testing equality of means across all 4 groups (3 treatment groups plus 1 control). For the non-dichotomous variables, the numbers in brackets represent the standard deviations.

Table 2. Summary Statistics for Study Participants

	Overall Mean	Control Mean	Item Difference	Money30 Difference	Money60 Difference	P-value of All Treatments=0
Age	35.0	34.4	0.57	0.81	1.14	0.89
	[14.2]	[13.6]				
Female	0.58	0.58	-0.07	0.05	0.02	0.15
University Affiliated	0.47	0.48	0.01	-0.02	-0.02	0.93
Student	0.43	0.46	-0.01	-0.03	-0.06	0.60
College degree or higher	0.88	0.9	-0.03	-0.03	-0.03	0.79
Exercise ≤ 1 day/week last year	0.43	0.36	0.08	0.12	0.09	0.14
No past exercise routine established	0.55	0.5	0.06	0.07	0.06	0.52
Expected avg weekly visits at this gym	3.1	3	0.04	-0.02	0.07	0.90
	[1.2]	[1.1]				
Perceived % chance of 9+ visits in 6 weeks	78.6	77.7	2.15	1.18	0.60	0.72
	[17.2]	[17.5]				
Number of Observations	690	176	156	176	182	

Table notes: Table presents information for new members who consented to participate in the study and were eligible for compensation. The overall mean column is the mean for the entire sample. The control mean column is the mean of the control group. The next three columns show the mean difference for the variable between the respective incentive groups and the control group. The p-value column displays the p-values testing equality of means across all 4 groups (3 treatment groups + 1 control). For the non-dichotomous variables, the numbers in brackets represent the standard deviations.

Table 3. OLS Regression Results of Treatments on Visit Measures

Panel A. Pooled analysis of all treatments vs control

	(1)	(2)	(3)	(4)	(5)	(6)
	9+ visits in	9+ visits in	Visits over	Visits over	Visits over	Visits over
Dependent variable:	1st 6 weeks	1st 6 weeks	1st 6 weeks	1st 6 weeks	weeks 7-12	weeks 7-12
incentive (pooled)	0.09**	0.10**	0.85	0.98*	0.18	0.45
	(0.04)	(0.04)	(0.59)	(0.58)	(0.59)	(0.58)
Controls	No	Yes	No	Yes	No	Yes
Observations	690	656	690	656	690	656
R-squared	0.01	0.08	0.003	0.11	0.0001	0.11
Control Mean of dep var	0.48	0.48	9.41	9.54	6.13	6.18

Note: Heteroskedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Controls in even number columns include age, gender, university affiliation, membership type, indicators for frequency of exercise in year before joining from the pre-survey, and an indicator for reporting no success in establishing an exercise routine in the past from the pre-survey. Observation counts in regression with controls are lower because 34 participants did not fully complete the pre-survey.

Panel B. Individual treatment estimates

	(1)	(2)	(3)	(4)	(5)	(6)
	9+ visits in	9+ visits in	Visits over	Visits over	Visits over	Visits over
Dependent variable:	1st 6 weeks	1st 6 weeks	1st 6 weeks	1st 6 weeks	weeks 7-12	weeks 7-12
item	0.09*	0.09*	1.34*	1.04	0.83	0.73
	(0.05)	(0.05)	(0.76)	(0.72)	(0.78)	(0.76)
money30	0.05	0.08	0.58	1.12	0.32	0.93
	(0.05)	(0.05)	(0.77)	(0.77)	(0.77)	(0.75)
money60	0.12**	0.12**	0.68	0.79	-0.50	-0.26
	(0.05)	(0.05)	(0.73)	(0.70)	(0.70)	(0.69)
Controls	No	Yes	No	Yes	No	Yes
Observations	690	656	690	656	690	656
R-squared	0.01	0.08	0.004	0.11	0.005	0.11
Control Mean of dep var	0.48	0.48	9.41	9.54	6.13	6.18

Note: Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Controls in even number columns include age, gender, university affiliation, membership type, indicators for frequency of exercise in year before joining from the pre-survey, and an indicator for reporting no success in establishing an exercise routine in the past from the pre-survey. Observation counts in regression with controls are lower because 34 participants did not fully complete the pre-survey.

Table 4. Means for Visit Measures by Treatment Offer

	Observations	9+ visits in 1st 6 weeks	Visits over 1st 6 weeks	Visits over weeks 7-12
Mean for all offered Control Group	207	0.45	9.08	5.96
		(0.03)	(0.47)	(0.47)
Mean for all offered incentives	628	0.52	9.65	5.94
		(0.02)	(0.30)	(0.29)
Difference with control mean		0.07	0.57	-0.02
p-value of difference from control		0.10	0.33	0.98
Mean for those offered Item	200	0.52	10.06	6.37
		(0.04)	(0.53)	(0.52)
Difference with control mean		0.07	0.98	0.41
p-value of difference from control		0.19	0.17	0.56
Mean for those offered Money30	214	0.50	9.54	6.12
		(0.03)	(0.52)	(0.52)
Difference with control mean		0.05	0.46	0.16
p-value of difference from control		0.35	0.51	0.75
Mean for those offered Money60	214	0.54	9.38	5.30
		(0.03)	(0.50)	(0.45)
Difference with control mean		0.09	0.30	-0.66
p-value of difference from control		0.07	0.66	0.31

Notes: Standard errors of means in parentheses. p-values are for a two sided t-test of equality of means. The number of observations differs across experimental groups because some individuals who were presented with gym enrollment packets either never actually joined the gym or merely wished to change their membership type. The latter were ineligible for the study because they were not new members.

Table 5. Treatment Heterogeneity by Measures of Past Exercise Patterns

Panel A. Split on self-reported frequency of exercise in the prior year

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	9+ visits in 1st 6 weeks		Visits over 1st 6 weeks		Visits over weeks 7-12	
Pre-survey variable split:	Past exercise ≤ 1 day/week	Past exercise > 1 day/week	Past exercise ≤ 1 day/week	Past exercise > 1 day/week	Past exercise ≤ 1 day/week	Past exercise > 1 day/week
incentive (pooled)	0.09	0.11*	1.07	0.88	0.83	0.05
	(0.07)	(0.06)	(0.71)	(0.83)	(0.66)	(0.85)
Additional controls	No	No	No	No	No	No
Observations	285	372	285	372	285	372
Control Mean of dep var	0.38	0.54	7.05	10.88	3.48	7.76

Note: Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Panel B. Split on self-reported success in establishing exercise routine in the past

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	9+ visits in 1st 6 weeks		Visits over 1st 6 weeks		Visits over weeks 7-12	
Pre-survey variable split:	Struggle w/ routine	Past routine established	Struggle w/ routine	Past routine established	Struggle w/ routine	Past routine established
incentive (pooled)	0.07	0.11*	1.01	0.45	0.53	-0.14
	(0.06)	(0.06)	(0.72)	(0.93)	(0.65)	(1.00)
Additional controls	No	No	No	No	No	No
Observations	361	296	361	296	361	296
Control Mean of dep var	0.42	0.55	7.88	11.19	4.42	7.95

Note: Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Indicator for struggle with routine in the past is set to 1 for those who selected in the pre-survey one of the following statements as the best fit for their past experience: "I have never tried to establish an exercise routine", "I have repeatedly tried to establish an exercise routine, but I have never been successful", or "I have at times established a regular exercise routine, but have been unable to stick to it for long periods". The other two options in the survey were: "Although I struggle with my commitment occasionally, I am usually able to keep up a regular exercise routine" and "I am a workout buff: I keep a regular exercise routine without much problem at all".

Figure 1a. Visit Rates Control vs Pooled Incentive Groups

Figure 1b. Visit Rates Control vs Incentive Groups

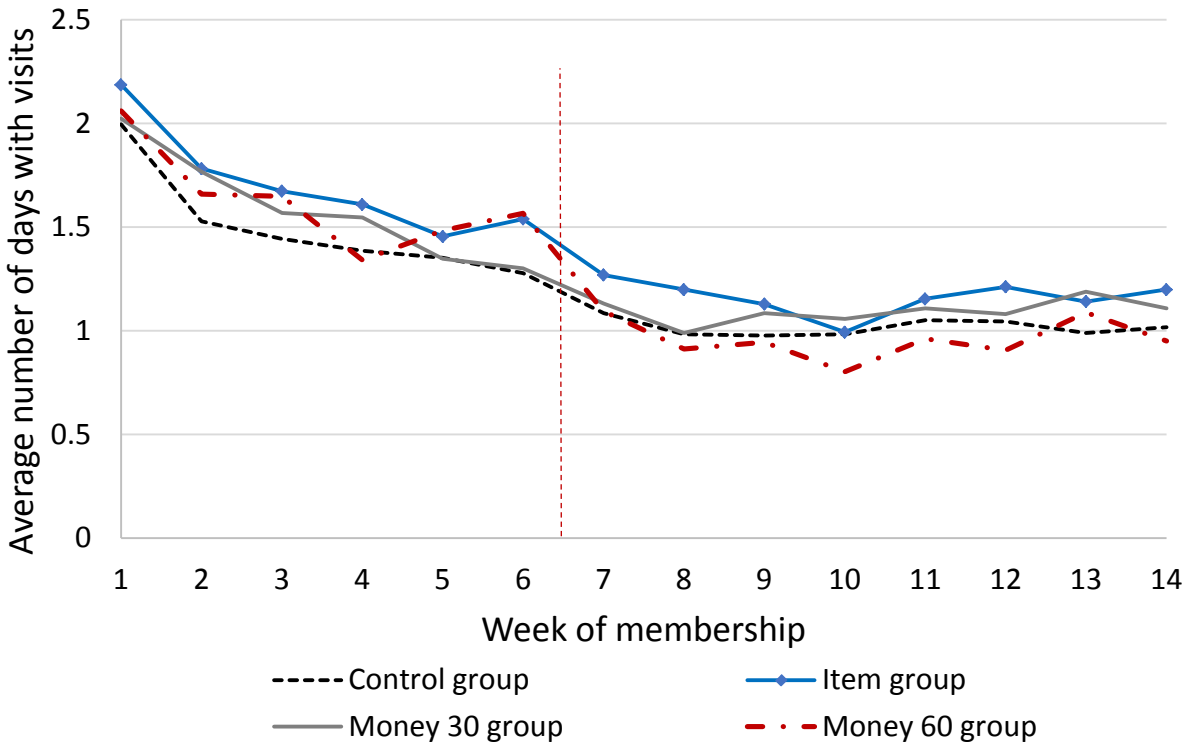

Figure 2. Histograms of Visits during First 6 Weeks of Membership by Treatment