Journal of Health Economics 58 (2018) 202–214

Contents lists available at ScienceDirect

Journal of Health Economics

journal homepage: www.elsevier.com/locate/econbase

Can financial incentives help people trying to establish new habits? Experimental evidence with new gym members[☆]

Mariana Carrera[a,*], Heather Royer[b], Mark Stehr[c], Justin Sydnor[d]

[a] Case Western Reserve University, United States
[b] University of California, Santa Barbara & IZA & NBER, United States
[c] Drexel University, United States
[d] University of Wisconsin, Madison & NBER, United States

ARTICLE INFO

Article history:
Received 28 June 2017
Received in revised form
29 December 2017
Accepted 15 February 2018
Available online 1 March 2018

JEL classification:
C93
D3
I12

Keywords:
Exercise
Incentives
Endowment effect
Overconfidence

ABSTRACT

Can financial incentives aid habit formation in people attempting to establish a positive health behavior? We provide evidence on this question from a randomized controlled trial of modest-sized incentives to attend the gym among new members of a fitness facility. Our experiment randomized 690 participants into a control group that received a $30 payment unconditionally or one of 3 incentive groups that received a payment for attending the gym at least 9 times over the first 6 weeks of membership. Two incentive treatment arms offered monetary payments of $30 and $60. The third incentive treatment, motivated by the endowment effect, offered a physical item worth $30. All three incentives had only small impacts on attendance during members' first 6 weeks and no effect on their post-incentive visit trajectories. We document substantial overconfidence among new members about their likely visits and discuss how overconfidence may undermine the effectiveness of incentive programs.

1. Introduction

Exercise is a prototypical example of a positive health behavior for which self-control problems have been argued 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 et al., 2015). Many also pay substantial fees over long periods for gym memberships that they do not use (Della Vigna and Malmendier, 2006).

A small set of randomized-controlled trials tests whether temporary financial incentives for exercising can increase exercise while the incentives are in place and ultimately help people establish lasting habits once the incentives are removed. In general, people seem to respond positively while incentives are active. In a number of cases incentives have been shown to create quite sizeable changes in behavior, though this is often in response to rather high monetary incentives with total possible rewards exceeding $100 (e.g., Charness and Gneezy, 2009; Royer et al., 2015; Carrera et al., 2017). Once the incentives are no longer in place, however, the evidence on habit formation is much more mixed, with lasting effects observed in a few cases (especially Charness and Gneezy, 2009) but overall tending to be modest or non-existent (Acland and Levy, 2015; Royer et al., 2015; Carrera et al., 2017; Rohde and Verbeke, 2017).

One potential reason this literature has shown relatively weak effects of incentives on helping people to establish exercise habits is that many prior studies have offered incentives to populations that are not already actively trying to change their behavior. The existing studies on exercise incentives typically recruit study participants from an underlying population, such as undergraduate students or employees, at a time that is unrelated to endogenous

[☆] We are thankful for grant support from the National Institute on Aging Roybal Center (grant P30AG034546, Dr. K. Volpp PI) 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 the Editor and two anonymous reviewers, Julien Mousquès, and various seminar and conference participants.

[*] Corresponding author.
E-mail address: mpc67@case.edu (M. Carrera).

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

https://doi.org/10.1016/j.jhealeco.2018.02.010
0167-6296/© 2018 Elsevier B.V. All rights reserved.

attempts at behavior change. Research in psychology suggests, however, that behavior change that leads to successful habit formation may require that people are motivated and prepared for making a change (Ajzen, 1985). Perhaps, then, timing incentive programs to coincide with moments when people have already taken the first step toward establishing a new habit, but have not yet settled into a new routine for that behavior, could be more effective. On the other hand, there are a number of potential forces that might make incentives ineffective for those who have initiated behavior change. For example, those who have started changing behavior may already be at peak motivation, leaving little room for incentive effects. People may also be overoptimistic or otherwise biased about their own behavior when they are first trying to change behavior and that could also interact negatively with some incentive designs.

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 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 (e.g., Della Vigna and Malmedier, 2006). These patterns of attendance are clear amongst our study population. New 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 on average 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 their exercise habits make this a valuable study population. We hypothesized that a temporary incentive during this initial period, when routines are adjusting and visits are declining, could help people make more regular visits during their initial membership phase and could in turn could generate lasting habit formation.

Our experiment randomized 690 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 prior to our intervention and resulted in a large mass of individuals whose expected behavior absent the incentive would put them somewhat near the target.[2] Although this target rate of attendance is slightly below that of previous work (e.g., Charness and Gneezy (2009) and Acland and Levy (2015) incentivized 2 visits per week on average over a 4-week period), our intention was to increase visit rates in the lower half of the distribution. From a health perspective, the marginal benefits of additional weekly exercise are likely largest among those who are exercising least.

The incentive arms consisted of two monetary incentives and one non-monetary (material) incentive. The monetary incentives were either $30 or $60 for reaching the 9-visit target, both paid in the form of an Amazon.com gift card. Subjects in the third incentive arm earned a specific but subject-chosen item sold by Amazon.com worth approximately $30. This item-based incentive was inspired by research on the endowment effect (Kahneman et al., 1990). We hypothesized that selecting a specific item at the outset might create a sense of ownership for that item so that not achieving the target visit rate might feel like a “loss” of the item. If this sense of ownership could be established with an item incentive, it might make an item incentive more powerful for loss-averse people 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 not effective at helping people to increase their exercise frequency. 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 the incentive period. We find that the item incentive induced slightly more visits than the equivalent-valued monetary incentive, but the differences are small and not statistically significant. In heterogeneity analysis, low exercisers and those who struggle with establishing exercise habits in the past had the largest increases in total visits in response to the incentives. Yet even for these groups, the effects are minor and not statistically significant. Overall, we conclude that the provision of modest additional financial incentives only marginally changed the behavior of new gym members.

Relative to prior work, our study offers at least four new contributions. We discuss related literature in more detail in the next section, but outline our contributions here. First, the primary contribution comes from testing the effect of incentives offered to the unique group of new gym members who are attempting to establish a habit of using the gym. We find that an additional incentive provided during this initial habit-formation period did not meaningfully improve the trajectory of habit formation. Second, we test the effect of relatively modest financial rewards (e.g., $30 or $60) on motivating exercise behavior. Modest-sized incentives (as opposed to the high-powered incentives tested in prior work) are relevant when considering broad and easily scalable interventions. Importantly, our design randomized the size of these incentives within the same incentive design, which is rare within the literature on tests of incentives for health-behavior change. As such, we can comment on the elasticity of response with respect to incentive size. We see little evidence of increased effects for the $60 incentive relative to the $30 incentive. Third, by testing the item incentive (which had some potential to invoke the endowment effect) relative to equivalent-valued monetary incentives, we add to a small literature exploring whether incentives that incorporate behavioral insights can be more powerful without additional cost (e.g., Patel et al., 2016a,b; Carrera et al., 2017). In our case, we find little support for this approach as a way of improving the power of the incentive program. Finally, and somewhat more subtly, we embedded our incentive offer into the standard enrollment procedures for new members at the gym, which allowed us to test the program on the full subpopulation of interest. This contrasts with most of the literature on exercise incentives, which have typically recruited study populations using surveys or other opt-in procedures and then randomized within that self-selected study group. Moving away from opt-in samples toward treatment offers randomized across an entire population of interest is challenging but important for understanding how interventions may scale, because those who opt into participating in studies may be more responsive to incentive programs than the broader population.

In the concluding section of the paper we discuss some potential implications of our findings and possible reasons that the incentives were ineffective in this setting. Prior studies of exercise incentives that did not target people who were starting to establish their exercise habit have found sizeable responses while incentives were in place using similar incentive designs to ours but with larger incentive payments (e.g., Charness and Gneezy, 2009; Acland and Levy, 2015). The modest-sized incentives in our study could be the reason for the smaller response. Yet while much larger incentives

[2] We used data on gym members who joined prior to our intervention to determine that 9 was the median number of visits in the first 6 weeks and that the distribution of visit counts was single-peaked. The distribution of visits among our control group subjects is similar.

might have had a stronger effect, we find little elasticity to incentive amounts between the $30 and $60 incentives. A more compelling possibility to us is that the threshold nature of the incentive program may be ineffective for those who are starting new exercise routines because it interacts negatively with overconfidence. We document that new gym members appear overconfident about how often they will visit the gym in the absence of an incentive. Our incentivized subjects, then, might have been overconfident about their baseline likelihood of hitting the incentive threshold at the start of their memberships, which might have weakened the power of the incentive to motivate behavior during the first few weeks of the membership. We discuss some ways in which future studies might better structure incentive programs to bolster endogenous attempts at habit formation in recognition of the potential for over-optimism in these populations.

2. Related literature

In this section we briefly review the existing literature on financial incentives for health-behavior change with an emphasis on where our study design offers new contributions. We focus mostly on randomized experiments testing incentives for physical activity and gym attendance but also touch on selected literature on incentives for weight-loss and smoking cessation as well.

Our work is closely related to a number of randomized experiments testing the provision of incentives for exercise (Courneya et al., 1997; Finkelstein et al., 2008; Charness and Gneezy, 2009; Babcock and Hartman, 2010; Pope and Harvey-Berino, 2013; Hunter et al., 2013; Pope and Harvey, 2014; Acland and Levy, 2015; Babcock et al., 2015; Royer et al., 2015; Patel et al., 2016a,b; Rhode and Verbeke, 2017; Carrera et al., 2017). Like our study, the majority of these studies incentivized attendance at a fitness facility, with the exceptions being three studies that incentivized physical activity through pedometers or other movement trackers (Finkelstein et al., 2008; Hunter et al., 2013; Patel et al., 2016a,b). The population bases for these studies differ, with some focusing on undergraduate students, others on employee populations, and one on a generic overweight adult population. The commonality in each case, however, is that subjects were recruited into the study at a moment in time that was not generally related to their own endogenous attempts to change behavior. This is the key contrast with our study, which recruited members of a fitness facility at the time when they initially joined. While there is obvious value in the existing literature that recruits from non-targeted populations, our study's focus on new members who are in the process of establishing their behavioral patterns allows us to bring an important new contribution to the literature.

While the literature on incentives for health-behavior change in other settings does not have an exact analogue to our focus on new members, it points to some reasons to believe that timing incentives to coincide with periods of high intrinsic motivation may matter. For example, studies in psychology have used survey instruments to measure such motivation, finding that it correlates positively with both short-run and long-run outcomes in weight loss programs (Williams et al., 1996) and diabetes treatment (Williams et al., 1998). Motivation also fluctuates over time, with recent research finding that many people start new exercise routines and make other life changes on salient dates, such as birthdays (Dai et al., 2014). Meta-analytic evidence suggests that incentives for tobacco cessation have larger effects on pregnant smokers than on broader smoking populations (Cahill et al., 2015), suggesting that incentives might work particularly well as complements to other motivations for behavior change.

We drew on the existing literature on exercise incentives when designing the structure and stakes for the incentive programs in this study. First, we used a threshold-target design, in which subjects had to attend the gym a minimum number of times in a fixed period to earn the incentive. This is by far the most common design in the literature, as all but three of the studies referenced in the second paragraph in this section used a threshold design. There are benefits and possible limitations to the use of threshold designs as opposed to alternatives such as per-visit incentives.[3] The literature has not yet rigorously tested the benefits of threshold-target versus per-visit incentives and our budget for this study was too limited to include additional treatment arms testing this difference. We see this as an important area for future research.

In setting the stakes for this program, we aimed to provide a moderate incentive level that was plausibly big enough to affect behavior but also potentially scalable. The variation in incentive designs across the existing literature makes it somewhat difficult to concretely compare the stakes across studies, but one rough measure is to consider the earnings a person who maximizes their earnings in the study receives per episode of exercise. Using this measure, our study provides incentives in the middle of the range of the existing literature at $3.33 and $6.67 per visit for the $30 and $60 incentive treatments respectively. Five previous studies used high-powered incentives of $10 or more per visit to the gym and in each case there was a large behavioral response to the incentives while they were in place (Charness and Gneezy, 2009; Babcock and Hartman, 2010; Acland and Levy, 2015; Royer et al., 2015; Carrera et al., 2017). Four previous studies tested incentives equating to less than $2 per episode of exercise and three of the four (including all of the studies with large sample sizes) reported small to non-existent behavioral effects (Finkelstein et al., 2008; Hunter et al., 2013; Patel et al., 2016a,b; Rohde and Verbeke, 2017). The three studies with middle-ground incentives of $4-$7.50 per action were mixed, with one showing no effect and the others sizeable effects (Courneya et al., 1997; Babcock et al., 2010; Pope and Harvey-Berino, 2013). This limited evidence suggested that moderate incentives in our incentive range might be effective. Importantly, much of our speculation about what size of incentives to adopt comes from cross-study comparisons, which are inherently flawed due to non-incentive differences across studies. Thus, randomizing the size of incentives within a study can lead to new insights about the elasticity of responses across incentive sizes. A few studies on weight-loss incentives have randomized the size of the incentive, often by sizeable amounts, and have actually tended to find weak effects of incentive size on treatment effects (Jeffrey et al., 1983; Augurzky et al., 2012; Paloyo et al., 2015). Our study is the first in the literature on exercise incentives to randomize the size of the incentive holding fixed other features of the incentive and hence provides unique evidence on the elasticity of response to incentive size.

Our third incentive arm testing the item incentive is related to a smaller segment of the literature on health-behavior incentives exploiting behavioral insights to increase the power of incentives without increasing the budget for incentives. These approaches are largely motivated by prospect theory (Kahneman and Tversky, 1979) and attempt to leverage loss aversion or probability weighting to increase response to incentives. Designs that incorporate a deposit portion that is forfeited by the subject if the behavioral target is not met are motivated in part by loss aversion and have been shown to successfully motivate weight loss (Volpp et al., 2008; John et al., 2011; Cawley and Price, 2013) and smoking cessation

[3] On the positive side, threshold incentives are fairly easy to describe to subjects, focus on meaningful levels of behavior change and reduce budgeting uncertainty for the incentive study. On the other hand, threshold incentives may provide little motivation for an individual who is far from earning the incentive and are sensitive to the target set in a way that per-visit incentives are not.

(Halpern et al., 2015), though only Halpern et al. (2015) directly compared a deposit design to a similar-sized "gain-only" design and found no difference in the treatment effect.

Arguably the most direct test of behaviorally-designed interventions comes from Patel et al. (2016a,b), who randomized subjects into a control group and three treatment groups: a "gain" group where a subject earns $1.40 if he completes a daily goal of 7000 steps, a "loss" group where a subject was given $42 in a month and $1.40 was taken away each day he missed the target of 7000 steps, and a lottery group in which a subject was entered into a lottery with an expected value of $1.40 if he made his target. Subjects in the "loss" group made the 7000 step goal a significantly greater fraction of the time than the control group while the "gain" and "lottery" treatments did not significantly differ from the control. On the other hand, List and Samek (2015) found that small incentives increased healthy food choices for school children but find little evidence framing of "gains" vs "losses" had a differential effect. Taken together these studies provide limited but suggestive evidence that leveraging loss aversion may be a way of increasing the power of incentives at low cost.[4]

One challenge to implementing "loss-framed" incentives, however, is that generating a sensation of "loss" in an incentive program can be difficult. Prior literature on the endowment effect demonstrates that invoking a sense of ownership over physical items can induce loss aversion (Kahneman et al., 1990). To the extent a self-selected item prize might generate a sense of ownership, this could provide a practical way of leveraging loss aversion without needing to create forms of deposit contracts or providing rewards that can later be taken back. Both of these incentive structures have obvious drawbacks – deposit contracts require individuals to trust the entity that offers such a contract and the taking back of rewards is complicated when individuals may have already consumed such rewards. However, it is unknown whether an item incentive can easily generate the sense of ownership required for the endowment effect to take hold. Our study design allows us to provide unique evidence on this front by directly comparing an item incentive to a monetary prize of equivalent value.

Finally, we note that nearly all studies of incentives for health behavior change have been conducted using an initial study-recruitment phase followed by randomization into incentive and control arms. This approach provides clean internal validity and avoids selection concerns. However, this approach may overstate the effect of incentives when they are offered at scale to broad populations if those who pay attention and respond to study invitations are more responsive to incentives. In this regard, our study is closer to studies by Cawley and Price (2013) on weight-loss incentives and Rohde and Verbeke (2017) on gym-attendance incentives, both of which studied incentive programs offered to a broad population rather than a pre-enrolled study sample. Our study is closest to Rohde and Verbeke (2017), who also study attendance incentives at a commercial fitness facility, but one important difference is that our design allows us to ensure that everyone in the population offered incentives was aware of the offer. In the Rohde and Verbeke study, in contrast, gym members were sent a letter informing them of the incentive offer but there was no way to track whether or not the letter was actually read. As we discuss below, however, these benefits of our design come with some challenges related to selection problems, which we believe we can effectively address but add some complication to the analysis.

3. Experimental design

3.1. Setting

Our experiment took place at a commercial gym consisting of roughly 3000 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% 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 gym is open 7 days per week from 5:30 a.m. to 12:30 a.m. on weekdays and 8 a.m.–10 p.m. on the weekend. Members have an ID card that is swiped by front desk personnel upon entry to the facility. These time-stamped entry records form the primary data for this study.[5]

We do not observe exercise behavior outside of this gym. Thus, any observed treatment effect could overstate changes in total exercise, if incentivized members substitute away from other forms of exercise to come to the gym.

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

3.2. 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.[6] We used a stratified randomization procedure where enrollment forms were sorted in stacks that alternated control and each of the three treatment assignments, with a separate stack for each of three membership types.[7] Gym staff simply used the enrollment packet on the top of the appropriate stack as new members joined the gym.

The enrollment forms included an IRB-approved consent form, short survey, and contact information sheet. Subjects had to con-

[4] There have also been tests in other literatures outside of health behaviors of whether loss aversion can be used to increase the power of incentives. Levitt et al. (2016) find no evidence that responses to incentives for students to perform well on test scores respond to loss framing. However, Fryer et al. (2012) report a strong effect of loss-framed incentives for test scores when they are provided to teachers.

[5] 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.

[6] 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."

[7] 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.

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

	Overall Mean	Control Mean	Item Difference	Money 30 Difference	Money 60 Difference	P-value of All Treatments = 0
Participation Rate	0.83	0.85	−0.07	−0.03	0.00	0.22
Age	35.3 [14.6]	35.1 [14.5]	−0.05	0.00	0.62	0.96
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 pool of new members who were invited to participate in the study and randomized into one of the four treatment arms. 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.

sent 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 to 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 reasonable. 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. We report the details and results for all treatment arms conducted in the study. These groups were as follows:

(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) and thus avoided selection concerns.[8]

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.[9]

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 on Amazon 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 et al., 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 an equivalent-valued monetary incentive. 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).

3.3. Summary statistics

Table 1 presents descriptive statistics for the full population of new members who joined the gym during our study period (N = 836), of whom 690 participated in the study. 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.[10]

Just over 80% of new members consented and were eligible to participate in the study. There are slight differences in participation rates across the experimental arms but we are unable to reject

[8] 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).

[9] 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.

[10] 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.

Table 2
Summary Statistics for Study Participants.

	Overall Mean	Control Mean	Item Difference	Money 30 Difference	Money 60 Difference	P-value of All Treatments=0
Age	35.0 [14.2]	34.4 [13.6]	0.57	0.81	1.14	0.89
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 $\leq$1day/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
Planned avg weekly visits at this gym	3.1 [1.2]	3 [1.1]	0.04	−0.02	0.07	0.90
Perceived% chance of 9+ visits in 6 weeks	78.6 [17.2]	77.7 [17.5]	2.15	1.18	0.60	0.72
Number of Observations	690	176	156	176	182	

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.

that the participation probability is the same across the treatments. Approximately half of the new members are associated with the nearby university and nearly 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.[11] 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.[12] The majority (55%) of subjects chose one of these three unsuccessful routine 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 planned to attend regularly. On average, new members reported plans to visit the gym 3 times per week. Interestingly, this number 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.[13] Again, interestingly, this was similar for those in the treatment and control groups.

4. Results

4.1. 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. Fig. 1ashows 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 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 reported planning to make 3 visits per week, on average, suggesting that without added incentives, they attended about one third as often as they had planned 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

[11] We note that self-reported measures of exercise are subject to overreporting bias (Sallis and Saelens, 2000). We use this measure only to define groups with above and below median self-reported past exercise for the purposes of heterogeneity analysis in Section 3.3.

[12] 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%).

[13] The survey question appears in the appendix. We reached the average of 78% reported in the text by assigning 10%, 30%, 50%, 70% and 90% to those responding 0–20%, 21–40%, 41–60%, 61–80% and 81–100% respectively.

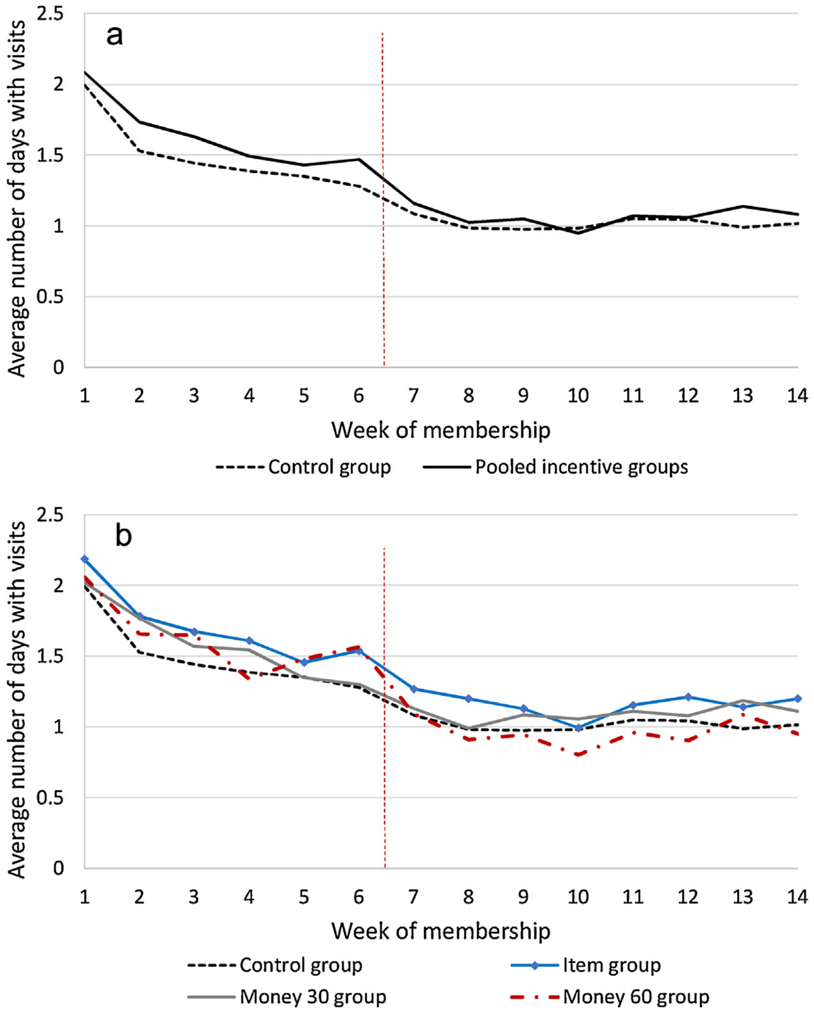

Fig. 1. (a) Visit Rates Control vs Pooled Incentive Groups. (b) Visit Rates Control vs Incentive Groups.

6 weeks of membership, consistent with the presence of incentives. The differences, however, were quite small and generally are not statistically significant (we provide formal tests in our regression analysis below). 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 6th 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 8 on, the average visit rates for the incentivized groups and control groups were very similar and hover around 1 visit per week.

Fig. 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 slight selection patterns identified in Tables 1 and 2 could be partly responsible for these results. As such, below we present regression results to quantify the difference in visit rates and to control for the small differences in observables between subjects in the different treatments.

It is not obvious that the average number of visits measure examined in Fig. 1a and b is the most relevant outcome measure 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. Fig. 2 shows histograms with the number of visits top-coded at 24 (average of 4 per week) due to the long and sparse right tail in visit counts. This top-coding is done only for presentation purposes here and apart from this figure, there is no top-coding in any of the empirical analyses. 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 between 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

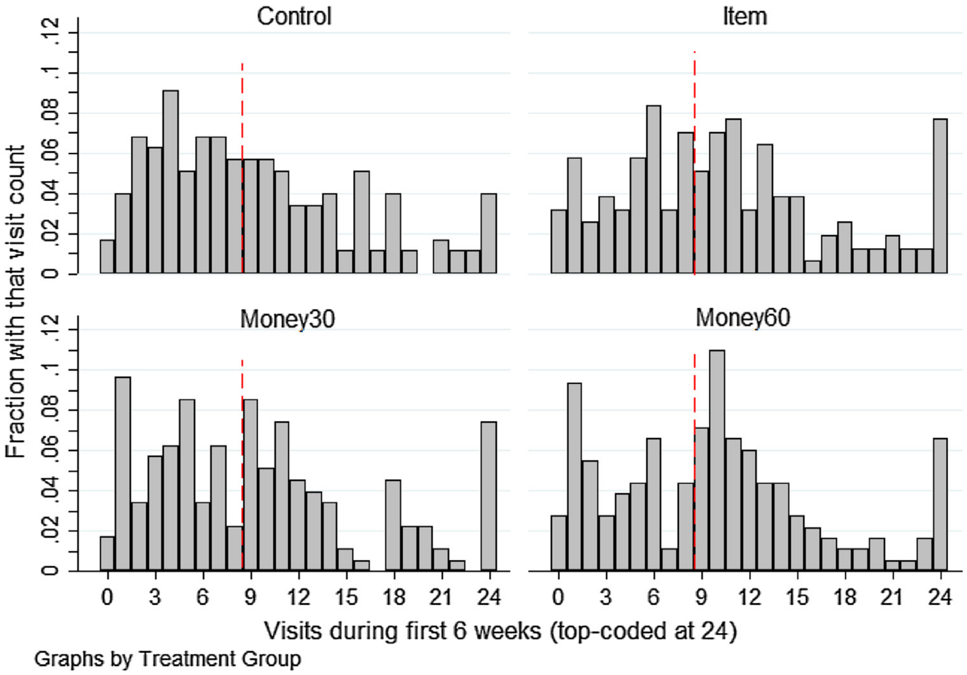

Fig. 2. Histograms of Visits during First 6 Weeks of Membership by Treatment.

visits in the first 6 weeks and hollowing out of the mass around 7 visits relative to that 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 Figs. 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 for those in the control group. 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; Royer et al., 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, which we include in even-numbered columns in both tables, has little impact on treatment effect point estimates. The matrix of controls include age in years, an indicator for being female, having a university affiliation, dummies for the membership type on which the randomization was stratified (e.g., student), and indicators for self-reported frequency of exercise in the year prior to joining the gym and self-reports of no success establishing an exercise routine in the past. Throughout we run ordinary least squares regressions with heteroscedasticity-consistent standard errors.[14]

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 were. This result is statistically significant (p = 0.02 with control variables included) and substantial relative to the control group's 48% probability of meeting the threshold. The increase in the average number of visits, however, is less pronounced. 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 1–6, in Fig. 1a. But, this difference is not statistically significant (p = 0.15). When controls are added, the estimated treatment effect increases slightly, to 0.98, but is only significant at the 10% level (p = 0.09). 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 Fig. 1a.[15]

[14] Results for the outcome of "9+ Visits" are robust to logit and probit specifications, and results for "Visits" are robust to Poisson regression and using ln(1+Visits) as the dependent variable.

[15] 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 between any 2 groups 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.

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)
Dependent variable:	9+ visits in 1 st 6 weeks	9+ visits in 1 st 6 weeks	Visits over 1 st 6 weeks	Visits over 1 st 6 weeks	Visits over weeks 7–12	Visits over 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

Panel B. Individual treatment estimates

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	9+ visits in 1 st 6 weeks	9+ visits in 1 st 6 weeks	Visits over 1 st 6 weeks	Visits over 1 st 6 weeks	Visits over weeks 7–12	Visits over 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

Notes: 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.

Panel B presents the same regression estimates for each treatment group separately.[16] Of the three incentive groups, Money 60 had the largest and most significant increase in the probability of meeting the 9 visit threshold, a 12 percentage point increase (p = 0.02 with controls). 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 disappointing average reflects the "hollowing-out" patterns seen in Fig. 2. Compared to the distribution of visits in the control group, the $60 incentive group 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 Item treatment was motivated by 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 (p = 0.09). The magnitude of the Item treatment effect on 9+ visits is not negligible – nearly a 20% increase in the probability of attending 9 or more days. We do not find any strong evidence to support the idea that the item incentive was more powerful. This lack of effect does not provide evidence against loss aversion or the endowment effect more generally, but suggests that those forces were not sufficiently activated or sufficiently strong in this item intervention to have a meaningful effect.

Because the point estimates suggest much smaller effects on average visits than we had anticipated, we are underpowered to precisely estimate these effects and differentiate effects between treatment arms. Our sample size—approximately 175 participants per treatment arm—enables us to detect differences in average visits over the first six weeks of 2 or more.[17] This minimum detectable difference is equivalent to 1/3 of a visit per week or an increase of approximately 20% relative to the control group. This effect would be slightly less than 30% of the effect size found by Charness and Gneezy (2009) during their incentive period. Since our incentives were at least 30% of the value of their $100 incentive, this was a reasonable expected effect size, ex ante. Also, if the $60 and $30 gift cards had effect sizes equal to 60% and 30% of Charness and Gneezy's $100 incentive effect, we would have been powered to detect the difference between those two treatment arms. In actuality, however, our results suggest that all three treatment arms have relatively small effects on average visit rates. Given the size suggested by their point estimates, we would need a far larger sample to detect those effects with 80% power. For example, to detect a difference of 1 visit over 6 weeks between two treatment arms, we would need 670 subjects in each treatment arm.

4.2. 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

[16] Again, results for the outcome of "9+ Visits" are robust to logit and probit specifications, and results for "Visits" are robust to Poisson regression and using ln(1+Visits) as the dependent variable.

[17] For all power calculations, we used 80% and 5% as the power and significance thresholds, respectively, and we used the standard deviation of visits over the first six week in our control group.

Table 4
Means for Visit Measures by Treatment Offer.

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 from 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 from 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 from 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 from 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 two sided *t*-tests 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.

hypothesis that treatment status had no effect on participation rates. Also it is worth recalling that the overall participation rate was high at 83%. 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 percentage point 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.

4.3. 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 too ambitious for some, 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.[18] 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.[19] 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.

Recall that our randomization was stratified across three types of members, who use different registration forms and face differently structured membership fees. We include controls for each membership type in all regressions that use controls. However, a referee raised the possibility that graduate students and wellness program participants, who do not pay standard monthly membership fees, may be less committed to starting a gym-going habit and thus, less responsive to our incentives. We present results run sep-

[18] Again, we caution that these self-reported values of past exercise behavior are likely inflated due to social desirability bias.

[19] This could also reflect the fact that these members might respond to the incentives by substituting away from exercise elsewhere to exercise at this gym. The scope for such substitution is naturally more limited among those reporting less ex-ante exercise.

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 1 st 6 weeks		Visits over 1 st 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

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

Notes: 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”.

Table 6
Treatment Heterogeneity by Membership Type.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent variable:	9+ visits in 1st 6 weeks			Visits over 1st 6 weeks			Visits over weeks 7–12		
Membership type:	Regular member	Graduate student	Wellness program	Regular member	Graduate student	Wellness program	Regular member	Graduate student	Wellness program
incentive (pooled)	0.13*	0.08	0.08	2.07**	1.07	−0.98	0.51	1.12	−0.82
	(0.07)	(0.07)	(0.09)	(0.83)	(0.96)	(1.36)	(0.99)	(0.86)	(1.25)
Controls	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	254	251	151	254	251	151	254	251	151
R-squared	0.10	0.03	0.09	0.18	0.05	0.15	0.14	0.08	0.14
Control Mean of dep var	0.54	0.40	0.53	10.06	8.58	10.34	7.22	5.07	6.42

Notes: Heteroscedasticity-robust standard errors in parentheses. *** p < 0.01, ** p < 0.05, * p < 0.1. See notes to Table 3 for a full description of control variables. Graduate students have a subsidized membership fee included by default with their tuition and fees. Members of a health insurer's wellness program are also able to obtain heavily subsidized memberships. Regular members pay an initiation fee and a monthly membership fee.

arately across the three membership types in Table 6. Although this analysis is based on a comparison across pre-specified randomized strata, we did not pre-specify or anticipate analyzing differences across these groups, so we advise some caution in interpreting the results.

Looking at Table 6, it does appear that the treatment effect of the pooled incentives is larger and more statistically significant among regular members than in the other two groups. This could be because people paying higher membership fees have either a greater motivation to earn something “back” (i.e., an income effect) or a stronger desire to establish a habit of regular attendance (i.e., a selection effect). Future work would be necessary to verify that these groups of members respond differently and to test possible explanations for such differences. We note, however, that even the larger effect we estimate on average visits for the group of regular members, an increase of approximately 1/3 of a visit per week, did not lead to increased exercise in the post-incentive period.

5. Discussion and conclusion

We conclude that the provision of moderately-sized financial incentives did not help new gym members establish more frequent gym habits. In this concluding section we discuss some potential implications of these results. Our evidence does not provide support for the hypothesis that timing incentives to align with endogenous attempts at habit formation is an effective strategy. It may be that additional incentives are simply ineffective for people who are starting a new gym membership. However, our study is only one data point and does not rule out that the possibility that alternative incentive schemes may lead to more promising results.

One question raised by our results is whether the small effects of our incentives are related to the size of the incentive. As discussed in Section 2, the incentive stakes in this experiment are smaller than those in previous studies that have documented substantial effects while incentives were in place. For example, our $60 treatment offered stakes around half the size of those in Charness and Gneezy (2009).[20] Also, our “monetary” treatments offered Amazon gift cards rather than cash, which might make them slightly less valuable to some participants, though surveys of our participants

[20] 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.

show that most are Amazon.com shoppers.[21] Ultimately, there is no way to rule out that larger incentives, on the order of those used in Charness and Gneezy might have induced stronger treatment effects. However, we note that we see little elasticity of response to incentive size when comparing the results of the $30 and $60 monetary treatment. We find little support for the idea that the exact incentive size is important within the range of moderate-sized incentives we test, which is an important class of incentives for scalable interventions.

Another possibility is that the limited effects of the incentive are related to the threshold-nature of the incentive design. As we noted in Section 2, most of the prior studies of exercise incentives had used similar threshold designs that required people to make a minimum number of visits over a pre-set time period to earn a reward. In particular, both Charness and Gneezy (2009) and Acland and Levy (2015) documented substantial average response to threshold-based incentives while they were in place and even saw some lasting effects on attendance once the incentives were removed. So, ex ante, there was no reason to expect that a threshold incentive would be ineffective for our study population. However, ex-post we believe there is a reasonable possibility that the threshold-nature of our incentive program may have interacted negatively with over-optimism among the new-member population. New members are extremely overoptimistic about how often they will visit the gym. According to our survey, 95% of participants planned to visit the gym more than once per week on average, but the share of participants who did so was 63% in the first month and dropped to 34% in the third month. Study participants also reported at the outset that they believed they would reach the 9-visit target with high probability. For example, the control group reported an average likelihood of reaching that visit rate of 78%. They were strongly over-confident in these beliefs on average, however, as only 48% of the control group actually achieved the 9-visit target. This overconfidence may be problematic for a threshold incentive design if it causes people to pay less attention to the incentive because they (wrongly) believe they will earn it easily. Speaking somewhat to that possibility is the fact that the three incentivized treatment groups did not report higher self-assessed likelihoods of reaching the 9-visit target at the outset than the control group even though they were aware of their incentive program. A related possibility is that during the incentive period, subjects might have overestimated how many visits they previously made, making them overconfident about having met the threshold. The effects of overconfidence are less likely to be problematic in prior studies that recruited from populations that were not already engaged in trying to change behavior. For those populations, they have a past history of attendance to predict their likelihood of success and thus, are better able to determine whether behaviors need to be changed to achieve their attendance target.

Given our findings, a fruitful area of new research may be to investigate other incentive programs that address the issue of over-optimism more successfully. For example, the response to per-activity incentives (e.g., per-visit incentives to the gym) should not affected as much by over-optimism about future behavior. However, the problem with per-visit incentives is that a substantial portion of the cost of such programs goes to paying the incentives to high-use members. For example, in our control group we observe that 16% reach the 9 visit threshold as early as their third week, but 14% do not come at all in weeks 2–4 and 28% average less than one visit per week in the first month. Another approach that could overcome over-optimism and has less of this problem of rewarding high-use members would be to set shorter durations for more frequent threshold incentives, such as rewards for each week in which the person made 1 or 2 visits.

More generally, future studies likely could benefit from exploring incentive payout frequency among those starting exercise routines. It is possible that the six week delay to receive the incentive was too far in the future to provide adequate motivation during the initial weeks of membership. Providing more immediate incentives during the first few weeks of membership or potentially timing the incentive program to begin in the second or third week as visit rates start to drop could be useful avenues for future research.

Ultimately, we believe these results suggest that simply timing an incentive program to coincide with endogenous attempts at habit formation 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, tempering 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 productive 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.

[21] 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.

References

Acland, Dan, Levy, Matthew R., 2015. Naiveté, projection bias, and habit formation in gym attendance. Manage. Sci. 61 (1), 146–160.

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

Augurzky, Boris, Bauer, Thomas K., Reichert, Arndt R., Schmidt, Christoph M., Tauchmann, Harald, 2012. Does Money Burn Fat? Evidence from a Randomized Experiment. IZA Discussion Paper No. 6888 (SSRN) https://ssrn.com/abstract=2158298.

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

Babcock, Philip, Bedard, Kelly, Charness, Gary, Hartman, John, Royer, Heather, 2015. Letting down the team? Social effects of team incentives. J. Eur. Econ. Assoc. 13 (5), 841–870.

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

Cawley, John, Price, Joshua A., 2013. A case study of a workplace wellness program that offers financial incentives for weight loss. J. Health Econ. 32 (5), 794–803.

Charness, Gary, Gneezy, Uri, 2009. Incentives to exercise. Econometrica 77 (3), 909–931.

Courneya, K.S., Estabrooks, P.A., Nigg, C.R., 1997. A simple reinforcement strategy for increasing attendance at a fitness facility. Health Educ. Behav. 24 (6), 708–715.

Dai, Hengchen, Milkman, Katherine L., Riis, Jason, 2014. The fresh start effect: temporal landmarks motivate aspirational behavior. Manage. Sci. 60 (10), 2563–2582.

Della Vigna, Stefano, Malmendier, Ulrike, 2006. Paying not to go to the gym. Am. Econ. Rev. 96 (3), 694–719.

Finkelstein, Eric A., Brown, D.S., Brown, D.R., Buchner, D.M., 2008. A randomized study of financial incentives to increase physical activity among sedentary older adults. Prev. Med. 47 (2), 182–187.

Fryer, Roland G., Levitt, Steven D., List, John A., Sadoff, Sally, 2012. Enhancing the Efficacy of Teacher Incentives Through Loss Aversion: A Field Experiment National Bureau of Economics Working Paper No. 201477.

Halpern, Scott, French, Benjamin, Small, Dylan S., et al., 2015. Randomized trial of four financial-incentive programs for smoking cessation. New Engl. J. Med. 372 (22), 2108–2117.

Hunter, Ruth F., Tully, Mark A., Davis, Michael, Stevenson, Michael, Frank, Kee., 2013. Physical activity loyalty cards for behavioral change: a quasi-experimental study. Am. J. Prev. Med. 45 (1), 56–63.

Jeffrey, Robert W., Gerber, Wendy M., Rosenthal, Barbara S., Lindquist, Ruth A., 1983. Monetary contracts in weight control: effectiveness of group and individual contracts of varying size. J. Consult. Clin. Psychol. 51 (2), 242–248.

John, Leslie K., Loewenstein, George, Troxel, Andrea B., Norton, Laurie, Fassbender, Jennifer E., Volpp, Kevin G., 2011. Financial incentives for extended weight loss: a randomized, controlled trial. J. Gen. Intern. Med. 26 (6), 621–626.

Kahneman, Daniel, Knetsch, Jack L., Thaler, Richard H., 1990. Experimental tests of the endowment effect and the coase theorem. J. Political Econ. 98 (6), 1325–1348.

Cahill, Kate, Hartmann-Boyce, Jamie, Perera, Rafael, 2015. Incentives for smoking cessation. Cochrane Database Syst. Rev., Art. No.: C D004307.

Levitt, Steven, List, John A., Neckermann, Susanne, Sadoff, Sally, 2016. The behavioralist goes to school: leveraging behavioral economics to improve educational performance. Am. Econ. J.: Econ. Policy 8 (4), 183–219.

List, John A., Samek, Anya, 2015. The behavioralist as nutritionist: leveraging behavioral economics to improve child food choice and consumption. J. Health Econ. 39, 135–146.

Paloyo, Alfredo, Reichert, Arndt R., Reuss-Borst, Monika, Tauchmann, Harald, 2015. Who responds to financial incnetives for weight loss? evidence from a randomized controlled trial. Soc. Sci. Med. 145, 44–52.

Patel, Mitesh S., Asch, David A., Rosin, Roy, Small, Dylan S., Bellamy, Scarlett L., Heuer, Jack, Sproat, Susan, et al., 2016a. Framing financial incentives to increase physical activity among overweight and obese adults. Ann. Intern. Med. 164 (6), 385–394.

Patel, Mitesh S., Asch, David A., Troxel, Andrea B., Fletcher, Michele, Osman-Koss, Rosemary, Brady, Jennifer, Wesby, Lisa, et al., 2016b. Premium-based financial incentives did not promote workplace weight loss in a 2013–15 study. Health Aff. (Millwood) 35 (1), 71–79.

Pope, Lizzy, Harvey, Jean, 2014. The impact of incentives on intrinsic and extrinsic motives for fitness-center attendance in college first-year students. Am. J. Health Promot. 29 (3), 192–199.

Pope, Lizzy, Harvey-Berino, Jean, 2013. Burn and earn: a randomized controlled trial incentivizing exercise during fall semester for college first-year students. Prev. Med. 56, 197–201.

Rohde, Kirsten I.M., Verbeke, Willem, 2017. We like to see you in the gym—a field experiment on financial incentives for short and long term gym attendance. J. Econ. Behav. Organ. 134, 388–407.

Royer, Heather, Stehr, Mark, Sydnor, Justin, 2015. Incentives, commitments, and habit formation in exercise: evidence from a field experiment with workers at a fortune-500 company. Am. Econ. J.: Appl. Econ. 7 (3), 51–84.

Sallis, J.F., Saelens, B.E., 2000. Assessment of physical activity by self-report: status, limitations, and future directions. Res. Q. Exerc. Sport 71 (Suppl. 2), 1–14.

Volpp, Kevin G., John, Leslie K., Troxel, Andrea B., Norton, Laurie, Fassbender, Jennifer, Loewenstein, George, 2008. Financial incentive–based approaches for weight loss: a randomized trial. J. Am. Med. Assoc. 300 (22), 2631–2637.

Williams, G.C., Grow, V.M., Freedman, Z.R., Ryan, R.M., Deci, E.L., 1996. Motivational predictors of weight loss and weight loss maintenance. J. Pers. Soc. Psychol. 70 (1), 115–126.

Williams, G.C., Freedman, Z.R., Deci, E.L., 1998. Supporting autonomy to motivate patients with diabetes for glucose control. Diabetes Care 21, 1644–1651.