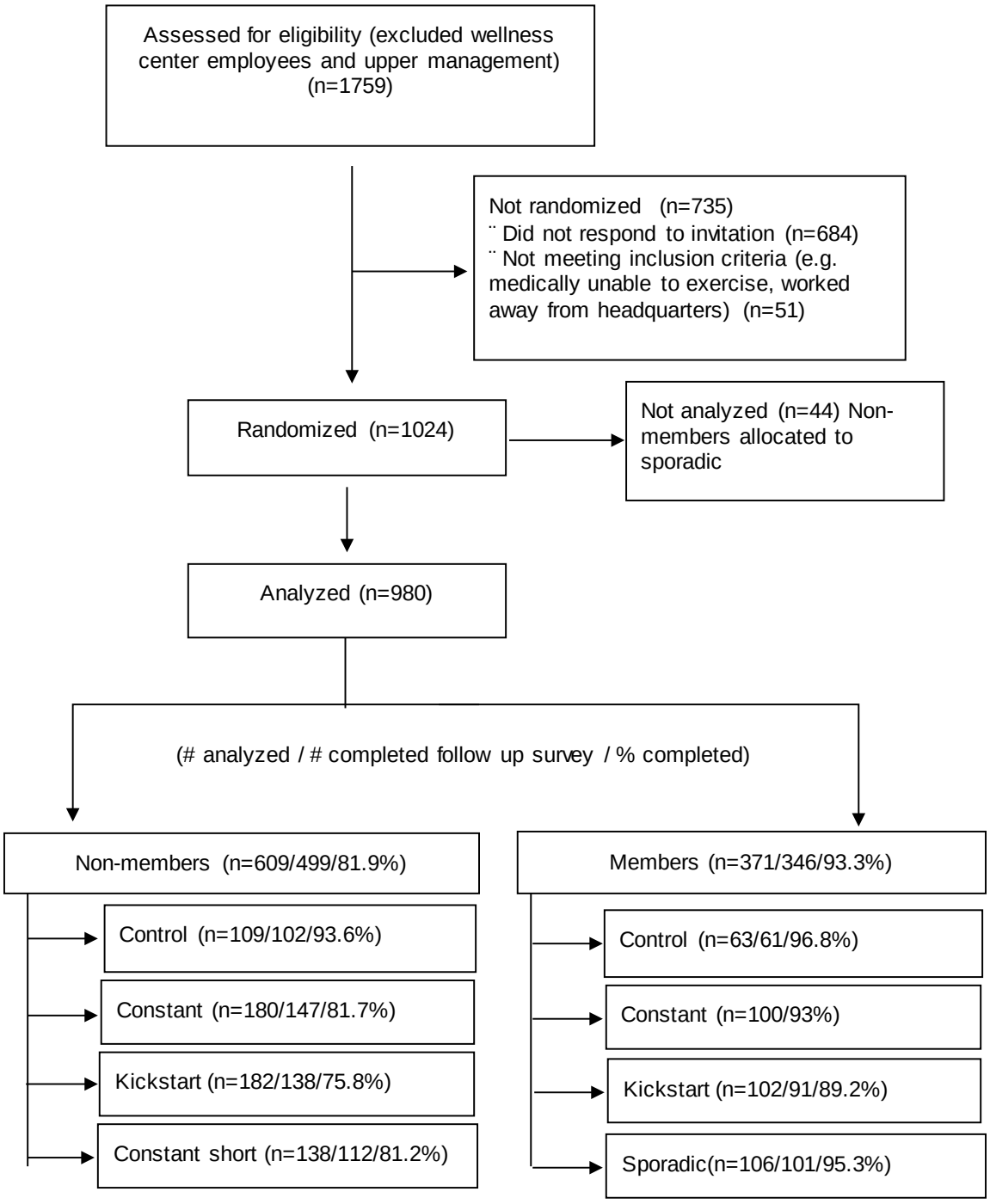
ONLINE APPENDIX

Appendix Figure A1. Experimental Participation Flowchart



A.1 Simulation of model with habit formation

Here we present simulations of the model presented in Section 2.3. We simulate the model to illustrate dynamics for a per-day incentive of $10 offered each day for an 8-week period (56 days). We assume that the future benefit generated by exercise is half as large as the incentive (i.e., $b = 5$). We assume the cost draws are normally distributed with mean $\mu = 5$ and $\sigma = 10$, so that the cost is lower than the future benefit half of the time and the incentive is equal to a standard deviation of the cost draw. Finally, we assume that habit stocks are governed by:

$$h_{t+1} = (1-\theta)h_t + \theta\left[a_t\bar{h} + (1-a_t)\underline{h}\right],$$

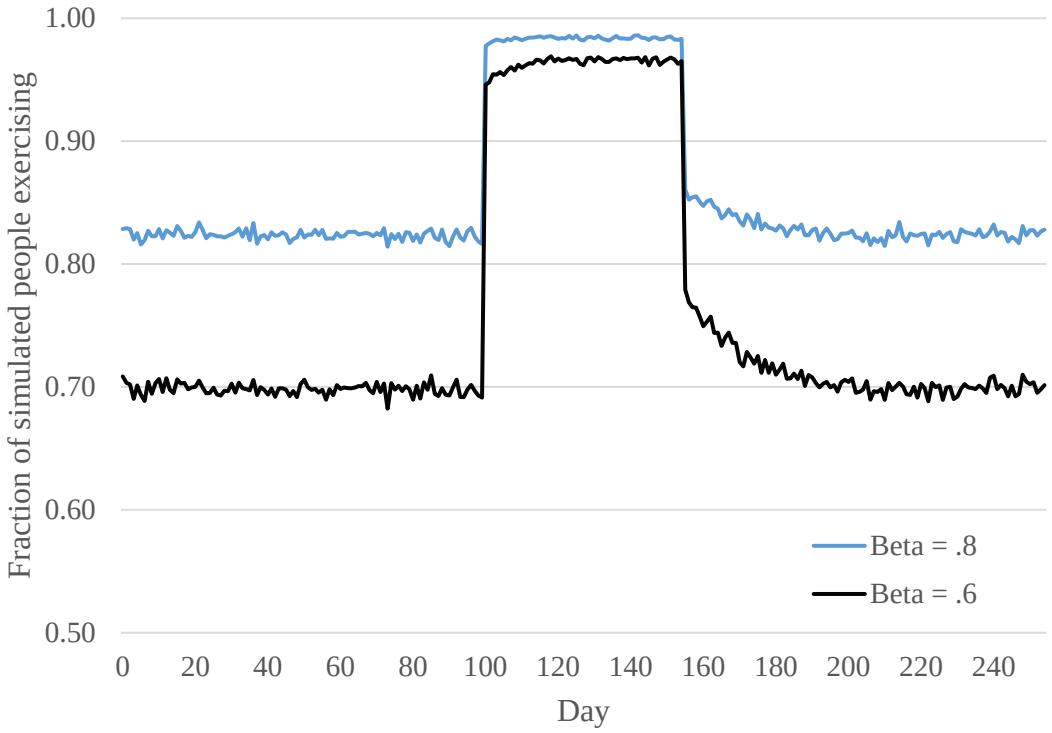
where θ is a parameter governing the speed of habit adjustment.[1] We set $\theta = 0.1$ for this example simulation so that habit stocks adjust 10% of the way toward the fully habituated and unhabituated levels depending on whether the person exercises or not. We set the fully habituated and unhabituated habit stocks to $\bar{h} = 5$ and $\underline{h} = -5$, which is equivalent to half a standard deviation in the daily cost distribution. These levels of modest habit stocks create a situation where habit stocks are not enough by themselves to overcome all possible cost draws and hence there is a chance of exercise when fully unhabituated and a chance of not exercising when fully habituated.

We simulated exercise patterns for 10,000 people who all begin at the fully unhabituated habit stock. After a long period of time, to allow behavior to reach an equilibrium, the simulation introduces the 8-week $10/day incentive as a surprise, though the individuals know the structure of the incentive (i.e., that it ends in 8 weeks) once the program begins. Appendix Figure A2 shows the fraction of simulated individuals exercising each day over time, starting at a point in the simulation far enough after the initial period that an equilibrium has been reached. We plot these dynamics for two levels of present-bias, $\beta = 0.8$ and 0.6. The incentive program significantly increases the fraction of people exercising each day, with a larger effect on those who are more present-biased since they begin from a lower equilibrium rate. After the incentive ends, exercise

[1] This is a general formulation for habit-stock adjustment that allows for a range of flexible habit-formation processes. For example, it embeds the auto-regressive process motivated by the Becker-Murphy (1988) framework used in Hussam et al., (2017). In that framework $h_t = \gamma h_{t-1} + a_t$, which can be produced in our framework by setting $\theta = 1-\gamma$, $\bar{h} = \frac{1}{1-\gamma}$, and $\underline{h}$=0.

rates remain elevated relative to the pre-incentive equlibrium due to the raised habit stock, but fall back toward the pre-incentive equilibrium over time.

Appendix Figure A2. Simulated exercise dynamics demonstrating "unsustained" habits



We also can use the same simulation structure and parameters to provide an example comparing a constant continuous incentive to a version where the incentive is turned on and off every other week. Specifically, Figure A3 shows the simulated fraction of people exercising for $\beta = 0.8$ under both a constant \$10/day incentive for 4 weeks and an alternative that offers \$10/day every other week for 8 weeks (i.e., weeks 2, 4, 6, and 8). The key point to notice in the figure is that the troughs of exercise during the non-incentivized weeks for the periodic incentive are all above the pre-incentive equilibrium. In the first non-incentivized week this is all due to forward-looking behavior by the agent who is recognizing that investing in exercise that week will have a dividend the following week when incentives are turned on because improving the habit stock will increase the likelihood of earning the incentive (see Hussam et al., 2017 for a discussion of this type of "rational addiction" process for health behaviors). In subsequent un-incentivized weeks visit rates are elevated due to a combination of that effect and the raised habit stock inherited from

the prior incentivized week. The constant incentive also yields elevated visit rates for the initial period after the incentive ends as the elevated habit slowly dissipates back to the baseline equilibrium. In this example, the overall average visit rate across the eight-week period is slightly higher for the periodic incentive than the constant incentive (91.34% vs. 91.28%).

Appendix Figure A3. Example simulated exercise dynamics for constant vs periodic incentive

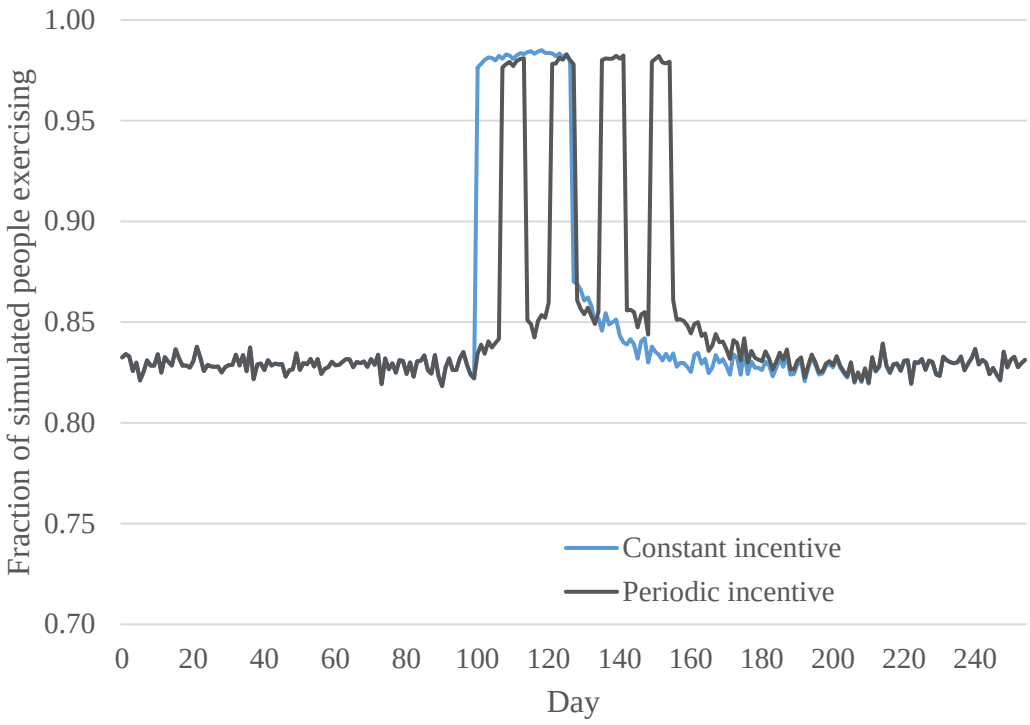
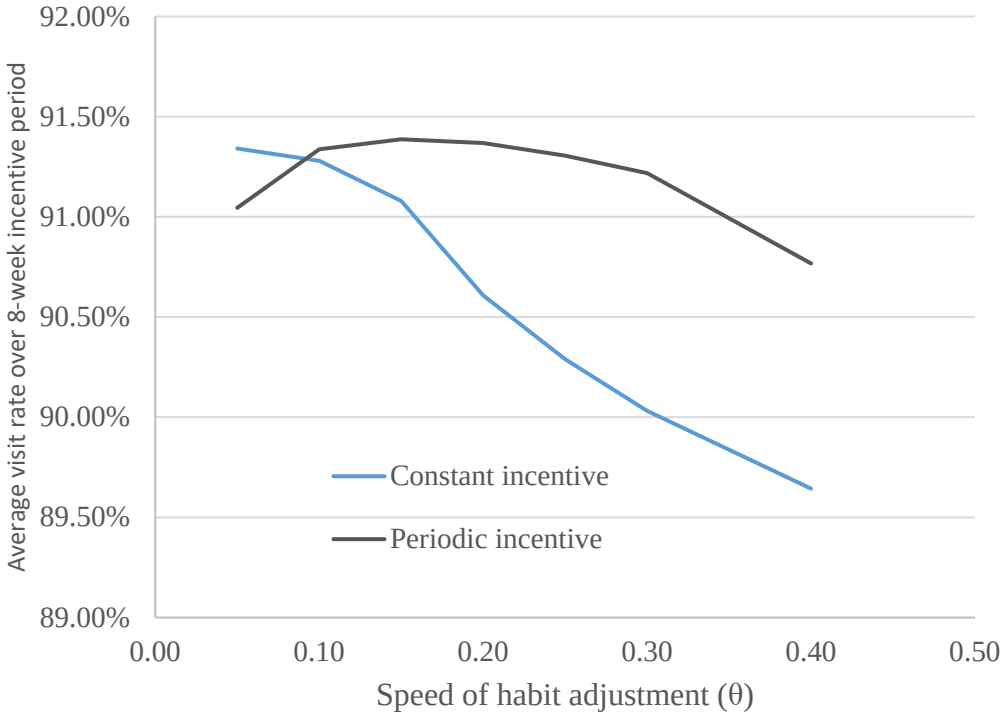


Figure A4. shows these average visit rates over the 8-week period as above for different speeds of habit adjustment (θ). For slower-building habits the average visit rate is higher for the constant incentive, but as the speed of habit adjustment rises the benefits of the periodic incentive get larger. We note that in all of these cases the average visit rate differences are small, but other parameter combinations at times yield more sizeable differences.

Appendix Figure A4. Comparison of average visits over 8 weeks for 4-week constant incentive vs. 8-week periodic incentive for different habit-adjustment speeds



A.2 Substitution Analysis

In this section of the appendix, we provide further details of the substitution analysis described in the main text. Our substitution analysis takes two steps: 1) assess how well self-reported measures work in calculating the treatment effects on company gym exercise and 2) then based on the findings of 1), proceed to estimate the treatment effects on overall exercise.

A slight complication arises when comparing treatment effects estimated for computerized logins with those estimated for total days of exercise from the survey. The computerized records should have little or no measurement error whereas the overall exercise measures are self-reported, and thus may be prone to recall error. To assess the validity of the self-reported measure, we estimate and then compare treatment effects from the computerized records with those from the

self-reported attendance. If those treatment effects are similar, then it may be reasonable to believe that the treatment effects on overall exercise days reported in the survey are also reliable.

Appendix Table A1 presents the effects of the incentive treatments on attendance from computerized company gym swipes, self-reported company gym days, and self-reported total exercise days separately for members and non-members. Note we combine all treatments (i.e., *Treated* is a binary indicator equal to 1 if one belongs to one of the incentivized groups and 0 otherwise) to increase precision in this analysis.

For members, the self-reported company gym attendance effects are significantly smaller than the company gym swipes effects. Thus, for this group, using the self-reported overall exercise days is unlikely to be fruitful because their self-reported company gym attendance is likely unreliable, and thus, their overall exercise measure is also likely subject to reporting biases. For members, each visit may be less salient and thus, their self-reported measures of exercise are more prone to measurement error. But, we worry less about substitution for this group because they are already members of the gym. According to the follow up survey, for those who were already members, 61% of total exercise days were company gym days, so the primary location of exercise for this group is the company gym.

For the non-members, the two estimates measuring company gym attendance effects are quite similar, 0.218 for company gym logins versus 0.186 for self-reported company gym days. Thus, the size of the total exercise days effect relative to the company gym days effect may be a reasonable measure of substitution. If we take the ratio of coefficients in columns (5) and (6), we calculate that 79% of the gym days induced by the incentives represented new exercise.[2] These results are similar to those in Royer, Stehr and Sydnor (2015) who followed a similar methodology and found that, again depending on the measure of company gym visits, either 63% (from self-reports) or 84% (from logins) of incentive-induced exercise at a gym represented new exercise for those who were not previously members of the gym. In summary, our results are consistent with the notion that for non-members, the incentive programs induced substantial new exercise.

[2] If we were to instead rely on the ratio of the coefficients in columns (4) and (6), this figure would fall to 67%.

A.3 Understanding the Quantity and Type of Exercise Induced by the Incentives

In this study, subjects in the treatment groups received incentives if the electronic records at the company gym showed that they made the required visits. Although this outcome measure has been used in the literature, it faces several shortcomings as a measure of new exercise. First, subjects may "game the system" by logging in and out of the gym without exercising and for the sole purpose of earning incentives. Second, as described in the body of the paper, subjects may substitute exercise at the company gym for other forms of exercise so that only a fraction of the exercise at the company gym is in fact new exercise. We address these concerns in two ways. The first relies on surveying subjects about their exercise behavior at and away from the company gym while the second relies on analyzing health outcomes that are plausibly affected by changes in exercise.

Our follow-up survey asked subjects how many minutes they spent exercising at the company gym, at other gyms and at places other than a gym during the calendar week before they took the survey. Controlling for exercise at these other venues, a regression of total minutes spent exercising on visits to the company gym shows that each visit to the company gym was associated with 45 additional minutes of exercise ($p = 0.01$). This relationship between minutes and visits was largely driven by exercise the subjects characterized as moderate to vigorous cardio and strength training rather than light cardio suggesting that the exercise was, in fact, substantive.

We also can test whether subjects in our treatment groups are reporting different levels of exercise per trip to the company gym than non-incentivized participants in the control group. To test this, we included in the regression described above a control group dummy and an interaction term between the control group dummy and the number of visits to the company gym. The coefficient on this interaction term was very small (-1.05 minutes) and statistically insignificant ($p = 0.91$), so we fail to reject that the relationship between visits to the company gym and minutes of exercise is the same for the control group and the treatment groups.

A.4 Estimating the Health Effects of the Incentives

As part of our study, we offered participants an additional incentive to provide health measurements at the time that they collected their survey payments.[3] While we anticipated ex-ante that we would be underpowered to detect health effects of the exercise induced by our intervention, the data we collected can aid future researchers in making power calculations for similar employee populations. For example, to detect treatment effects of 2 pounds or more (an effect that to us, seems reasonable), we would need to multiply our sample by 30. We present and discuss our health measure results, along with their caveats, in this section.

The ideal way to test for health effects of exercise induced by an intervention would be to collect all subjects' health measures before and after the intervention period, as done by Charness and Gneezy (2009) with a college student population. In corporate settings like ours, however, requiring multiple in–person visits for health measures could greatly reduce employees' interest in participating in any study or wellness program. Thus, for some of the health measures we collected, our analysis is limited to a comparison of ex-post measurements between treated and control groups, rather than a comparison of individual-level changes between pre- and post-intervention measurements. For these measures (blood pressure, pulse), any initial differences between the treatment and control groups cannot be observed, making our identification of a treatment effect rely on the assumption that initial (ex-ante levels) were equal on average. For weight and BMI, on the other hand, we can calculate individual-level *differences* between self-reported values provided in the initial survey and ex-post measurements.[4] For all comparisons, we pool together the different incentive programs in order to increase power, comparing them against the control group alone.

[3] We took advantage of the fact that all study participants had to visit the gym in person to pick up their earnings (including compensation for taking the survey as well as any earned incentives) at the end of the study, and asked them at this time to record their health measures (including weight) for an added payment of $25.

[4] While initial self-reported weight and height measures are likely prone to measurement error, most likely a systematic underreporting of weight, this should be orthogonal to the assignment of incentives given that the initial survey preceded subjects' assignment to incentive programs. It is possible, however, that self-reported weight in the follow up survey was influenced by the effectiveness of the treatment. For example, subjects in a more successful treatment might be pleased with the results and report lower weight than those in a less successful treatment. Such a reporting bias would lead to an upward bias in the treatment effect. The possibility of this type of bias, along with the possible bias resulting from self-selection into providing health measures, adds to our concerns that the observed changes in weight are not unbiased estimates of the treatment effects.

Regardless of whether the comparison between treated and control groups uses only ex-post values or individual-level differences, the analysis necessarily excludes all subjects who did not provide their health measures at the end of the 8-week study period. This raises the important concern of self-selection into the subsample providing health measures. We obtained ex-post health measures for 77% of the control group and 70% of the incentivized groups.[5] If employees who exercised more in response to the incentives were more likely to self-select into showing up and providing health measures, but there was no similar selection within the control group, this would bias us towards overestimating the treatment effect of the incentives on weight. Interestingly, there appears to be far less selection in our sample of gym members, of whom 88.9% of the control group and 86.7% of the treated participants gave health measures, as opposed to the non-members.

In Appendix Table A2, we report the difference-in-difference estimates of changes in weight caused by our intervention in three populations: all study participants, non-members and members, and show p-values based on the Wilcoxon rank-sum test of the difference-in-differences. In all populations, the incentivized groups appear to have a smaller increase in weight between the self-reported ex-ante measure and the objective ex-post measure, although the size and significance varies.[6] The difference-in-differences suggests that our incentives led to a statistically significant average weight reduction of 1.78 pounds among non-members, while the estimate for members was smaller (0.97 pounds) and statistically insignificant. Since the estimated treatment effects of our incentive programs on number of gym visits were much larger for members versus non-members, this contrast appears somewhat puzzling. In the absence of selection concerns, the larger impact on weight for non-members despite a smaller effect on gym visits could potentially be explained by stronger weight-loss induced by changes in exercise for those who had lower baseline levels of exercise. However, this pattern could easily be explained by different rates of

[5] Perhaps because they had fewer opportunities to earn money throughout the study, employees in the control group were also more likely than those receiving incentives to complete the follow-up survey (94.8% vs. 84.0%), more likely to pick up their earnings, and therefore more likely to provide their health measures at the time of earnings pickup.

[6] The fact that the "change" we measure is generally positive reflects individuals' tendencies to underreport their weight. If we assume that weight in the control group was constant between the time of the initial survey and the date when health measures were taken, our results suggest that individuals underreported their weight by an average of 7.4 pounds.

self-selection into providing health measures between members and non-members, described in the previous paragraph.

The self-selection concern is amplified by the fact that the estimated reductions in weight associated with receiving incentives are implausibly large considering the small number of gym visits (on average) induced by the incentives. For the pooled sample of members and nonmembers, the estimates suggest that one additional gym visit per week, over 8 weeks, is associated with weight loss of 4.53 pounds, implying that the average induced gym visit burned 0.57 pounds or 1,981 calories based on the commonly accepted "3,500 calories per pound" rule.[7] Based on the average body weight in our sample, burning 1,981 calories would require 2.4 hours of vigorous exercise such as running at 6mph or cycling 14-15.9 mph.[8] This is far more than the 45-minute average duration of a visit reported by participants (see Section Appendix A.3). Thus, we suspect that self-selection may be biasing our estimates of changes in weight.

For the remaining health measures, in which we only have ex-post measurements, we report averages for the treated and control groups in Appendix Table A3, again pooling all incentive programs together. Results are extremely noisy, with no t-tests detecting differences between the groups at conventional significance levels. Although insignificant, the differences indicate that the treated employees were 3.4 percentage points more likely to have high blood pressure readings and had slightly higher pulse readings (1.07 more beats per minute). This is likely related to the fact that within our sample providing health measures, the average self-reported weight was higher in the treated than control groups (180.6 vs. 176.3 pounds). However, to statistically detect this small treatment difference at a p-value of 0.05 we would need a sample approximately 10 times larger than what we have here in this study. For future studies interested in assessing the health effects, we suggest narrowing the number of treatments and increasing the sample size considerably to get more power for these health outcomes.

[7]JAMA Patient Page "Healthy Weight Loss" *JAMA.* 2014;312(9):974. doi:10.1001/jama.2014.10929

[8] Source: NutriStrategy. Calculations are based on research data from Medicine and Science in Sports and Exercise, the official journal of the American College of Sports Medicine. http://www.nutristrategy.com/caloriesburned.htm

Appendix Table A1. Substitution Analysis of In-Treatment Effects

	Members			Non-members		
	Company gym swipes	Company gym days from survey	Total exercise days from survey	Company gym swipes	Company gym days from survey	Total exercise days from survey
Treated	0.265	-0.03	-0.162	0.218***	0.186***	0.147
	(0.251)	(0.256)	(0.245)	(0.044)	(0.065)	(0.201)
Mean of dependent variable in control group in pre period	1.93	2.28	3.61	0.00	0.10	2.25
Observations	343	343	343	490	490	490

Table notes: Heteroskedasticty-robust standard errors in parentheses. *** p< 0.01, ** p < 0.05, * p < 0.10. Analysis is based on gym data and data from follow up survey. The dependent variable is days of exercise per week as measured by gym swipes or self-reported survey data. Treated includes treatment groups constant, kickstart and extended-sporadic for members and constant, kickstart and constant-short for non-members.

Appendix Table A2. Changes in Weight

	Control Group			Incentivized Groups			Difference-in-differences	
	N	Ex-ante (self-report)	Ex-post minus ex-ante	N	Ex-ante (self-report)	Ex-post minus ex-ante	$D_{incentivized} - D_{control}$	Change associated with an increase of one visit per week
Weight (lbs.)								
Full sample	130	176.3	7.45	560	180.6	6.05	-1.40*	-4.53
		[4.16]	[1.14]		[1.78]	[0.405]	[0.051]	
Non-members	75	181.4	7.50	297	183.8	5.72	-1.78**	-8.77
		[5.42]	[1.76]		[2.59]	[0.58]	[0.029]	
Members	55	169.4	7.39	263	176.8	6.42	-0.97	-2.24
		[6.43]	[1.24]		[2.38]	[0.56]	[0.65]	

Table notes: Participants gave self-reported weight in the initial survey ("Ex-ante") and provided objective weight measurements in Weeks 9-10 ("Ex-post"). The sample sizes shown are smaller than the full sample in the rest of the paper because we exclude participants who did not opt to provide health measures at the conclusion of the study. For the differences in differences column, p-values of a Wilcoxon rank-sum test are shown below the difference-in-differences. In the other columns, standard errors are shown below the mean values. In the last column, the difference-in-differences is rescaled based on the average increase in visits per week between control and incentivized groups in each of the three populations. *** p<0.01, ** p<0.05, * p<0.1.

Appendix Table A3. Other Health Measures

	Control Group		Incentivized Groups		Difference
	N	Mean [Std. Err.]	N	Mean [Std. Err.]	Incentivized mean - Control group mean
High blood pressure (0/1)					
Full sample	130	0.331	551	0.365	0.034
		[0.041]		[0.021]	[.047]
Non-members	75	0.373	291	0.423	0.050
		[0.056]		[0.029]	[.064]
Members	55	0.273	260	0.300	0.027
		[0.061]		[0.028]	[.068]
Members with <2 visits per week, ex-ante	30	0.267	135	0.274	0.007
		[0.082]		[0.035]	[.090]
Pulse (beats per minute)					
Full sample	129	73.24	551	74.31	1.07
		[1.07]		[.538]	[1.23]
Non-members	74	75.62	291	76.59	0.97
		[1.36]		[0.75]	[1.63]
Members	55	70.04	260	71.76	1.72
		[1.65]		[0.74]	[1.79]
Members with <2 visits per week, ex-ante	30	72.70	135	73.59	0.89
		[1.68]		[0.96]	[2.19]

Table notes: For this analysis, our sample is limited to participants who opted to provide their health measures, which were collected after the post-survey following the first 8 weeks of the study. Following the CDC definition, we classify an individual as having high blood pressure if their systolic reading is 140mmHg or higher or if their diastolic reading is 90mmHg or higher. Standard errors are shown below the mean values and mean differences.*** p<0.01, ** p<0.05, * p<0.1.