DS241 Final Project Report

Austin Ho, Vincent Qu, Joshua Shin, Jackie Wang

Abstract

At the start of the year, many people make New Year's resolutions in the hope of developing new habits. A popular example is the spike in gym membership sign-ups in January representing peoples' interests to exercise more. Based on Google Trends search data, the term "gym memberships" is searched nearly twice as much on New Year's Day as compared to the months leading up to it. However, another study published in ScienceDirect shows that many people who start new gym memberships at the beginning of the year do retain the habit across time, with a mean of 7.48 visits per month in the first month compared to 0.92 visits in the 12th month of the year. Our research study compares the habit development process for subjects when there is a daily or immediate reward system after completing a task compared to a lump sum reward system at the end of a time period if the subject completed all tasks over the study period. The results will show if and which reward system is more incentivizing to retain a new habit, and if the methodology of the reward system has any differing impact. This will demonstrate the effect that different reward systems can have on helping people develop new habits. Businesses can use the results of this study to increase their customer retention, consumers can use this to retain habits over time, and educators could use it to better incentivize their students.

Background

Forming new habits requires a complex mixture of motivation, discipline, and reward. Building a strong understanding of the mechanisms that foster new habits is critical and applicable to many fields, such as data science and economics. Rewards have usually served as the main motivator for behavioral change, although the timing and distribution of said rewards can have varied effects. Some strategies call for small, incremental rewards that create consistent incentives linked to the task assigned. Other strategies suggest creating a deferred rewards system, offering significant rewards at the end of the task period as the main motivation. This study seeks to empirically test both of these reward systems to determine which is more effective at promoting and maintaining new habits. The results from this study have potentially widespread applications to habit formation in the workplace, education sector, and personal life.

This experiment focuses on the incentive theory of motivation and explores what drives human behavior, and if certain interventions drive specific behaviors (Silverman, Jarvis, Jessel & Lopez, 2016). Incentive interventions are often used in the real world to lead humans to desired outcomes. Businesses use loyalty points to get customers to come back to the store while educators may award gold stars to students to promote good behavior. These practices stem from research suggesting that much of human behavior is influenced by its consequences (Cheney & Pierce, 2013). Through altering the specifics of the intervention, such as providing daily rewards versus a final lump sum, we expect differences in the habit-building behavior between the two groups that received a different reward system.

The specific concept we are investigating is whether altering the cadence of delivering rewards can affect the probability that a person will complete a routine activity given a set number of consecutive days. We found two similar studies that look at the effect of a reward system. One study looks at how the size and expectation of a reward can impact the amount of effort invested into a task (Frömer, Lin & Wolf, 2021). The study had three groups: one group was told they would receive a flat payment, another would receive rewards based on their performance, and the last group would be rewarded randomly. To measure the impact of these different reward systems, the researchers compared the speed that subjects in each group were able to complete the task correctly. Another study looked specifically at performance-based rewards at companies and how the design of this compensation impacted the employees' performance. Participants in the study were given timed activities to complete, with one group being rewarded with cash rewards while the other group received different tangible rewards. Impact of the reward system was determined by measuring the number of completed tasks within time limits set (Newman, Tafkov, Waddoups & Xiong, 2022). The results of these studies found varying degrees of task completion rates, and supports our conjecture that reward systems that offer up-front or immediate rewards see higher task completion rates than other reward systems. This also supports our intuition that people are impatient and are more motivated by tangible rewards immediately after task completion, as they have a clear goal to work towards that result in instant gratification.

Research Question

The role of the reward system in helping form habits is a highly studied and contested topic. One of the core discussions centers around what type of reward structure is the most effective in reinforcing habits. As such, this study will center around the following research question: *Does a daily reward system motivate people more to learn a new habit compared to a lump sum reward at the end of the specified habit-building time period?* Through altering the specific intervention, such as providing daily rewards versus a lump sum, we expect to see differences in the habit-building behavior between the treatment and control groups.

Hypothesis

Our hypothesis is that the treatment group (people who receive daily/immediate rewards) will perform better than the control group (people who receive a lump-sum payment for completion of all the tasks). Here, we define "performance" as the completion rate, or the total number of days completed, of our daily assigned tasks. We expect the outcome of the treatment effect to be positive. This hypothesis is based on the principle of immediate gratification, as we believe that the instant rewards following task completion will serve as strong motivation that reinforces habit creation. This reward system will provide participants with quick, tangible rewards while also strengthening their desire to continue doing the target task we assign. Furthermore, we suspect that this positive reinforcement will maintain high levels of motivation amongst treatment participants compared to the delayed rewards received by the control group.

Treatment

For our treatment, participants were incentivized with daily/immediate monetary payout upon completion of our designated task of 20 jumping jacks per day. For each day that the treatment group completes, they will receive a payout of $2. Conversely, the control group received a lump-sum reward of $10 only if they successfully completed all daily tasks over the designated time period of 5 consecutive days, with an average payout of $2 per day. The key distinction between the treatment and control groups lies in the reward system employed. Participants' progress throughout the experiment was monitored through video recordings submitted by participants showcasing task completion. We anticipate that the treatment group will exhibit higher rates of task completion via a positive treatment effect.

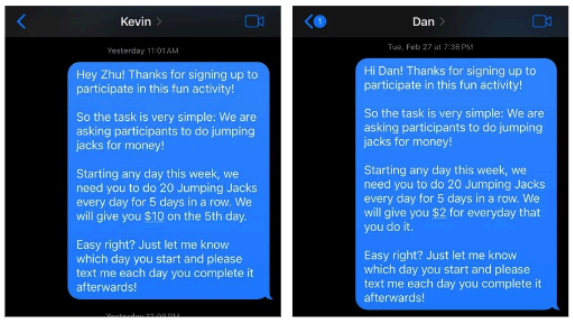


Figure 1: Participant Recruitment

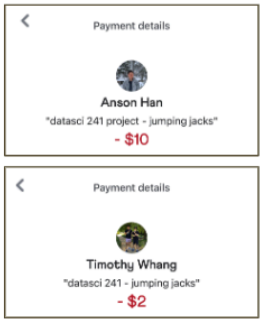


Figure 2: Payouts

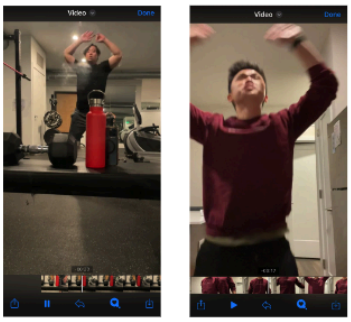


Figure 3: Video Evidence

Measurement Units

Direct recruitment from the researchers' social circles facilitated the formation of an attempted representative sample encompassing individuals from diverse ethnic backgrounds, gender identities, and location, aged between 18 and 60 years old. Our outreach efforts primarily leveraged phone calls, messaging platforms, and social media channels in order to optimize convenience of communication and help facilitate the seamless delivery of video recordings. Each researcher undertook the responsibility of contacting approximately 12-13 participants, a necessity dictated by our budgetary constraints amounting to $500 ($10*50 people = $500). Over the course of the experiment, we received low response rates from prospective participants, and increased the number of people contacted to about 100. Consequently, we aimed to enlist a total of 50 participants, with 25 assigned to treatment and 25 assigned to control. Additionally, we promptly addressed participant inquiries, meticulously tracked task completion rates daily, and diligently recorded any notable observations throughout the experiment. The measurement unit for the treatment variable was defined as the total number of days of task completion.

Randomization

Given our budgetary limitations, we aimed to recruit 50 participants for our study. However, due to a response and participation rate of about 50%, we expanded our outreach efforts, reaching out to roughly 100 prospective participants. Out of this pool, 52 individuals expressed interest and actively engaged in our study by initiating the daily tasks. Subsequently, we employed a fair coin flip to randomly assign each participant to either the treatment or control

group. The treatment group comprised 25 subjects, while the control group included 27 subjects. Utilizing this randomization method effectively ensured an approximately equal distribution between the treatment and control groups, with a split of 48% and 52%, respectively.

Flow Diagram

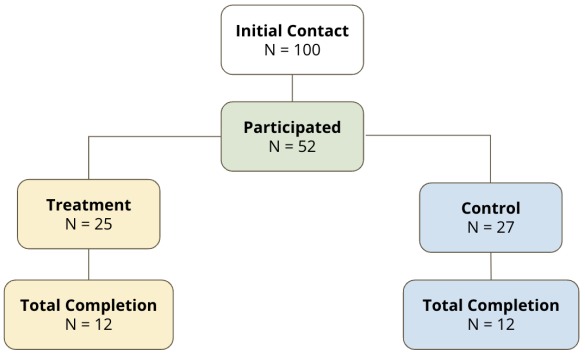


Figure 4: Flow Diagram of Participants

As mentioned above, we initially contacted around 100 people to recruit participants for our experiment, resulting in 52 people expressing interest and starting our study. 25 of these participants were sorted into the treatment group, and the remaining 27 were sorted into the control group. By the end of our study, there were 24 participants that completed all 5 days of the assigned tasks, with an even split of 12 people from treatment and 12 people from control.

Experimental Design

Our experiment was conducted as a randomized controlled trial, as we randomly assigned participants into either the treatment or control group using a fair coin flip. To minimize potential bias, the experiment was single-blind on the participants' side, as participants did not know whether they were sorted into the treatment or control groups. We had preordained outcome measures of testing for the total number of days of task completion. The experiment was a between-subjects design, as we compared the differences in outcomes between the treatment and control groups. We calculated the Average Treatment Effect (ATE) to compare differences in outcomes between the two groups. Although we attempted to collect data from a representative sample, this might be a potential limitation due to challenges in obtaining a truly representative sample within our social groups. Despite this, we diversified our participant pool by considering factors such as age, gender, occupation, location, and level of fitness.

Upon conclusion of the experiment, we recorded the total number of days of task completion for each group. Furthermore, we examined if there was a decline in the task completion rate for each consecutive day across both groups and within each group, along with identifying the proportion of participants who fully completed the study. This methodology was applied to both the treatment and control groups to determine if a daily reward system outperforms a lump sum reward system as an incentive for task completion. To ensure accuracy in assessing the degree of task completion, we utilized video evidence of the subjects performing

the tasks. As a result of this robust methodology, we believe that we can make a causal claim regarding the differences in incentive systems, on average, for this sample of participants.

Outcome Measures

Data was collected daily from participants to determine whether they had completed their assigned daily task. This information was stored in a spreadsheet for later analysis. To ensure the validity of the data, video evidence was submitted by the participants as proof of task completion. If a participant failed to complete the daily task in either treatment or control groups, they received a score of 0 for that day and were subsequently removed from the experiment.

The collected data consisted of column variables such as Treatment (with the control group coded as 0 and treatment group coded as 1), Day 1, Day 2, Day 3, Day 4, Day 5, and Total Days Completed. For each day, participants received a score of 1 if they completed the task and 0 otherwise. The Total Number of Days Completed variable was calculated as the sum of Days 1-5. As before, we measured the total number of days of task completion for each group to compare the effectiveness of daily reward systems versus lump sum reward systems as incentives for completing tasks, utilizing the Total Days Completed variable.

We also collected participants' demographic data which was included as additional covariates in our analysis to come. Furthermore, we examined if there was a decline in the task completion rate across days within both groups. This methodology was applied to both the treatment and control groups to assess the relative effectiveness of the two reward systems.

Analysis and Results

We first read in the CSV file where we gathered the original data recorded on the spreadsheet, and showed the below dataframe containing the first 6 rows of the data we collected from the study, with a total of 12 column variables.

	Age	Male	Caucasian	Asian	Bay	Treatment	Day1	Day2	Day3	Day4	Day5	TotalDaysCompleted
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
1	30	1	0	1	0	1	1	1	1	1	0	4
2	30	1	0	1	0	0	1	1	1	1	1	5
3	30	1	0	1	0	1	1	1	1	0	0	3
4	31	1	0	1	1	1	1	1	0	0	0	2
5	28	0	0	1	1	0	1	0	0	0	0	1
6	27	0	0	1	1	1	1	0	0	0	0	1

Figure 5: Original Dataframe Containing All Data Collected

In order to get a better understanding of the data, we ran some descriptive statistics formatted into a neat table using the "stargazer" package in R that shows us the count, mean, standard deviation, minimum, and maximum of all the covariates (Figure 6). We also produced a histogram of the outcome variable TotalDaysCompleted (Figure 7).

Statistic	N	Mean	St. Dev.	Min	Max
Age	52	24.635	5.784	20	60
Male	52	0.654	0.480	0	1
Caucasian	52	0.096	0.298	0	1
Asian	52	0.865	0.345	0	1
Bay	52	0.577	0.499	0	1
Treatment	52	0.481	0.505	0	1
Day1	52	1.000	0.000	1	1
Day2	52	0.788	0.412	0	1
Day3	52	0.635	0.486	0	1
Day4	52	0.577	0.499	0	1
Day5	52	0.462	0.503	0	1
TotalDaysCompleted	52	3.462	1.674	1	5

Figure 6: Stargazer Descriptive Statistics (All)

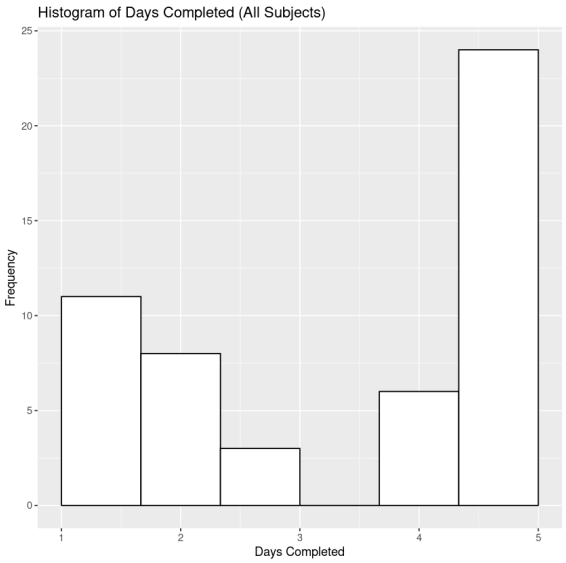


Figure 7: Histogram of Days Completed (All)

We then separated the original dataframe into two new dataframes: control group data (Figure 8) and treatment group data (Figure 9) below.

	Age	Male	Caucasian	Asian	Bay	Treatment	Day1	Day2	Day3	Day4	Day5	TotalDaysCompleted
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
1	30	1	0	1	0	0	1	1	1	1	1	5
2	28	0	0	1	1	0	1	0	0	0	0	1
3	23	1	0	1	0	0	1	1	1	1	1	5
4	30	1	0	1	0	0	1	1	1	1	1	5
5	22	1	0	1	0	0	1	1	1	1	1	5
6	23	1	0	1	0	0	1	0	0	0	0	1

Figure 8: Control Group Dataframe

	Age	Male	Caucasian	Asian	Bay	Treatment	Day1	Day2	Day3	Day4	Day5	TotalDaysCompleted
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
1	30	1	0	1	0	1	1	1	1	1	0	4
2	30	1	0	1	0	1	1	1	1	0	0	3
3	31	1	0	1	1	1	1	1	0	0	0	2
4	27	0	0	1	1	1	1	0	0	0	0	1
5	27	1	0	1	0	1	1	1	0	0	0	2
6	22	1	0	0	0	1	1	1	1	0	0	3

Figure 9: Treatment Group Dataframe

We, again, showed a descriptive statistics summary table (Figure 10) and the distribution for the TotalDaysCompleted outcome variable via a histogram (Figure 11) for the control dataset.

Statistic	N	Mean	St. Dev.	Min	Max
Age	27	25.370	7.556	21	60
Male	27	0.667	0.480	0	1
Caucasian	27	0.074	0.267	0	1
Asian	27	0.889	0.320	0	1
Bay	27	0.481	0.509	0	1
Treatment	27	0.000	0.000	0	0
Day1	27	1.000	0.000	1	1
Day2	27	0.741	0.447	0	1
Day3	27	0.556	0.506	0	1
Day4	27	0.556	0.506	0	1
Day5	27	0.444	0.506	0	1
TotalDaysCompleted	27	3.296	1.772	1	5

Figure 10: Stargazer Statistics (Control)

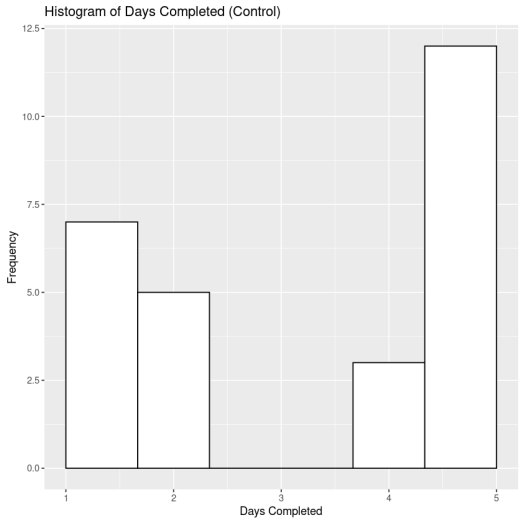


Figure 11: Histogram of Days Completed (Control)

The same process was repeated for the treatment dataset, with a descriptive statistics summary table (Figure 12) and histogram of the TotalDaysCompleted variable (Figure 13).

Statistic	N	Mean	St. Dev.	Min	Max
Age	25	23.840	2.824	20	31
Male	25	0.640	0.490	0	1
Caucasian	25	0.120	0.332	0	1
Asian	25	0.840	0.374	0	1
Bay	25	0.680	0.476	0	1
Treatment	25	1.000	0.000	1	1
Day1	25	1.000	0.000	1	1
Day2	25	0.840	0.374	0	1
Day3	25	0.720	0.458	0	1
Day4	25	0.600	0.500	0	1
Day5	25	0.480	0.510	0	1
TotalDaysCompleted	25	3.640	1.578	1	5

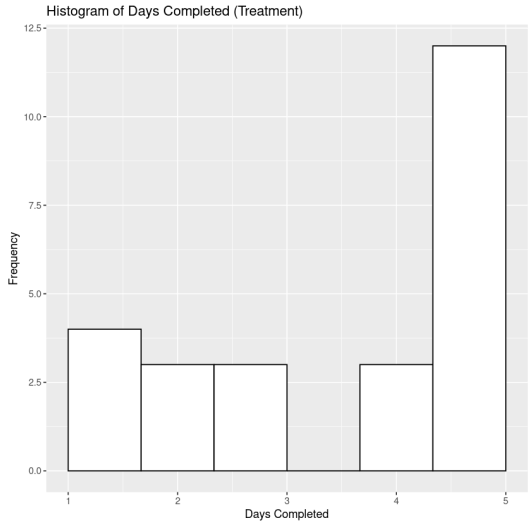


Figure 12: Stargazer Statistics (Treatment) **Figure 13:** Histogram of Days Completed (Treatment)

We arrive at the meat of the analysis, starting with a simple linear regression to calculate for the average treatment effect using the following formula: reg = lm(TotalDaysCompleted ~ Treatment, data = data). A well-formatted summary statistics table for the regression was output using the "stargazer" package (Figure 14), with additional information regarding residuals, t-values, and significance levels shown in the original regression summary output (Figure 15).

	Dependent variable:
	TotalDaysCompleted
Treatment	0.344
	(0.467)
Constant	3.296***
	(0.324)
Observations	52
R2	0.011
Adjusted R2	−0.009
Residual Std. Error	1.682 (df = 50)
F Statistic	0.542 (df = 1; 50)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
Call:
lm(formula = TotalDaysCompleted ~ Treatment, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6400 -1.6400  0.7037  1.3600  1.7037

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2963     0.3236  10.186 8.63e-14 ***
Treatment     0.3437     0.4667   0.736    0.465
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.682 on 50 degrees of freedom
Multiple R-squared:  0.01073,   Adjusted R-squared:  -0.009056
F-statistic: 0.5423 on 1 and 50 DF,  p-value: 0.4649
```

Figure 14: Stargazer Regression Summary **Figure 15:** Regression Summary

The above results show an average treatment effect of 0.3437, which means that the predicted number of total days completed in the study is, on average, 0.3437 days higher for the treatment group that received daily/immediate monetary rewards compared to the control group that received a lump-sum payout at the end of the study timeframe if they completed all days of the assigned tasks. This value was taken from the coefficient estimate on the "Treatment" variable regressed on the outcome variable of interest, TotalDaysCompleted. However, this result

is not statistically significant at any commonly used significance levels, as shown in the corresponding p-value of 0.465. To further solidify our regression results, we ran a coeftest in R (t-test of coefficients) to test for heteroskedastic-robust standard errors, in case there were potential issues of heteroskedasticity in our original model. This was done with the call: coeftest(reg, vcov = vcovHC(reg, type = 'HC0')), and the results are shown below (Figure 16).

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  3.29630    0.33463  9.8507 2.661e-13 ***
Treatment    0.34370    0.45562  0.7544    0.4542
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 16: T-test of Coefficients from Regression

The results from the coeftest are incredibly similar to the results from the regression summary table, with the coefficient estimate on “Treatment” staying roughly the same, and the results still being statistically insignificant. We also conducted a Welch two sample t-test in order to test our null and alternative hypotheses in a more robust manner. The null hypothesis is that the difference in means between the Treatment and Control groups is equal to zero, with the alternative hypothesis being that the difference in means between the Treatment and Control groups is greater than zero. The code is as follows, with the results of the t-test shown below in Figure 17: t.test(controldata$TotalDaysCompleted, treatmentdata$TotalDaysCompleted, alternative = c("greater"),mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95).

```
        Welch Two Sample t-test

data:  controldata$TotalDaysCompleted and treatmentdata$TotalDaysCompleted
t = -0.73974, df = 49.931, p-value = 0.7685
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -1.122401       Inf
sample estimates:
mean of x mean of y
 3.296296  3.640000
```

Figure 17: Welch Two Sample T-test of Regression

The results from the two sample t-test show that the 95% confidence interval is [-1.122401, Inf] and contains zero, meaning that the true value for the difference in means between the Treatment and Control groups could be zero. The associated t-value is statistically insignificant at all commonly used significance levels, and we conclude that we fail to reject the null hypothesis that the treatment has no effect on the outcome.

We then conducted several power analyses of our experiment under three different categories: pre-experiment assumptions, actual experimental outcomes, and an ideal experiment where we have at least 90% statistical power. The results are shown in Figures 18-20 below.

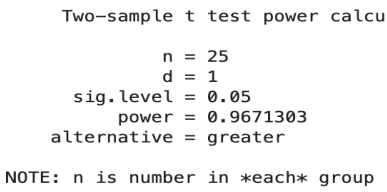


Figure 18: Pre-experiment Power

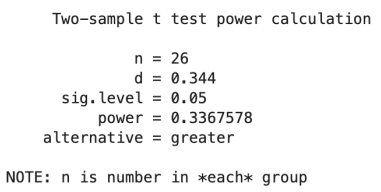


Figure 19: Experiment Power

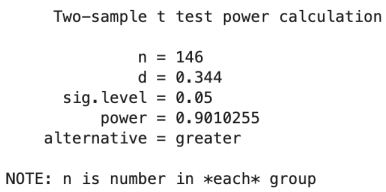


Figure 20: Ideal Power

In the pre-experiment case, we used the previous assumptions of how our experiment would look to calculate for power, with 25 people each in the Control and Treatment groups, an assumed effect size of 1 (meaning that we expected the Treatment group to have an average of 1 more day of task completion compared to the Control group), at a significance level of alpha=0.05, and with the alternative hypothesis being that the Treatment group would see a positive difference in means compared to the control group. The calculated power in this scenario is 0.9671303, meaning that we would have at least a 96.71303% chance of finding a statistically significant difference between the Treatment and Control groups when there exists one. In the experiment case, we used values that were actually obtained throughout the study, with roughly 26 people each in the Control and Treatment groups, a calculated effect size of 0.344 (meaning that the Treatment group was predicted to have an average of 0.344 more days of task completion compared to the Control group via the simple linear regression output), at a significance level of alpha=0.05, and with the alternative hypothesis being that the Treatment group would see a positive difference in means compared to the control group. The calculated power in this scenario is 0.3347578, which is unfortunately low. Finally, we tested an ideal situation in which we would be able to gather more participants in order to see a higher power of at least 90% using the same calculated effect size as produced in our experiment. We varied the value for the number of participants in each group until the power reached 90%, which resulted in 146 people each in the Control and Treatment groups, with the same calculated effect sizes, significance levels, and alternative hypothesis as in the experiment power calculation. The calculated power in this ideal scenario is 0.9010255, meaning that, if we wanted to have at least a 90% chance of finding a statistically significant difference between the Treatment and Control groups when there exists one at the effect size of 0.344 more TotalDaysCompleted for the Treatment group, we needed to recruit at least 292 participants, ceteris paribus.

In order to try to increase the statistical power of our analysis, as well as test for any other possible influences to our outcome variable, we decided to run a more complex regression with additional covariates using demographic data we collected on our participants. We included the variables Age (participants age in years), Male (indicator variable for whether the participant was male or female), Caucasian (indicator variable for whether the participant was Caucasian or not), Asian (indicator variable for whether the participant was Asian or not), and Bay (indicator variable for whether someone was from the Bay Area or not). If both variables Caucasian and Asian were zero, then the participant was Hispanic. The new regression was run with the following code: reg1 = lm(TotalDaysCompleted ~ Treatment + Age + Male + Caucasian + Asian

+ Bay, data = data), which was then formatted into a stargazer table (Figure 21) and produced the summary table with additional information (Figure 22).

```
================================================
                        Dependent variable:
                    ----------------------------
                           TotalDaysCompleted
------------------------------------------------
Treatment                         0.392
                                 (0.495)

Age                               0.043
                                 (0.044)

Male                              0.117
                                 (0.529)

Caucasian                         0.688
                                 (1.510)

Asian                            -0.118
                                 (1.305)

Bay                              -0.081
                                 (0.537)

Constant                          2.212
                                 (1.710)

------------------------------------------------
Observations                        52
R2                                0.055
Adjusted R2                      -0.071
Residual Std. Error         1.732 (df = 45)
F Statistic               0.437 (df = 6; 45)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

```
Call:
lm(formula = TotalDaysCompleted ~ Treatment + Age + Male + Caucasian +
    Asian + Bay, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-2.599 -1.796  0.482  1.489  2.034

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.21208    1.71037   1.293    0.202
Treatment    0.39190    0.49508   0.792    0.433
Age          0.04335    0.04379   0.990    0.327
Male         0.11651    0.52885   0.220    0.827
Caucasian    0.68846    1.51025   0.456    0.651
Asian       -0.11835    1.30528  -0.091    0.928
Bay         -0.08132    0.53738  -0.151    0.880

Residual standard error: 1.732 on 45 degrees of freedom
Multiple R-squared:  0.055,     Adjusted R-squared:  -0.071
F-statistic: 0.4365 on 6 and 45 DF,  p-value: 0.8505
```

Figure 21: Stargazer Regression (Covariates) **Figure 22:** Regression Summary (Covariates)

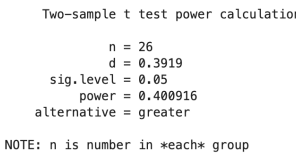
The above results for the regression with additional covariates show an average treatment effect of 0.3919, which means that the predicted number of total days completed in the study is, on average, 0.3919 days higher for the treatment group compared to the control group. However, this result is still not statistically significant at any commonly used significance levels, as shown in the corresponding p-value of roughly 0.433. Including the additional demographic covariates increased our calculated average treatment effect and lowered the associated p-value, but the results are not statistically significant enough to prove that there is indeed a non-zero treatment effect. To solidify our regression results, we again ran a coeftest producing heteroskedastic-robust standard errors, which was done with the call: coeftest(reg1, vcov = vcovHC(reg1, type = 'HC0')), and the results are shown in Figure 23 below.

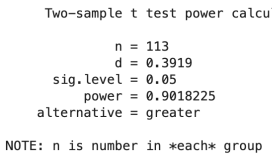
```
t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.212079   0.994523  2.2243  0.03119 *
Treatment    0.391898   0.461198  0.8497  0.39997
Age          0.043354   0.024636  1.7597  0.08525 .
Male         0.116506   0.497785  0.2340  0.81601
Caucasian    0.688464   0.857688  0.8027  0.42637
Asian       -0.118355   0.672482 -0.1760  0.86109
Bay         -0.081322   0.525543 -0.1547  0.87772
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

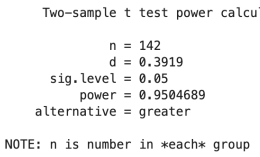
Figure 23: T-test of Coefficients (Covariates)

The results from the above coeftest are very similar to the results from the regression summary table for reg1, with the coefficient estimate on "Treatment" staying roughly the same, and the results still being statistically insignificant. We conducted another Welch two sample t-test to test our null and alternative hypotheses in a more robust manner, with the same null and alternative hypotheses as before, but the results remained the same as in the previous two sample t-test, in which we failed to reject the null hypothesis that the true difference in means between the Treatment and Control groups equals zero.

We re-conducted power analyses of our experiment using the effect size from the more complex regression (Reg1), and calculated three additional ideal scenarios with increasing power between 90%-99% shown below in Figures 24-27.







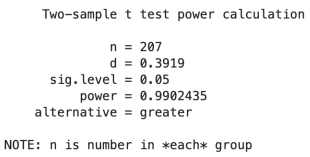


Figure 24: Reg1 Power **Figure 25:** 90% Power **Figure 26:** 95% Power **Figure 27:** 99% Power

While our statistical power increased with the addition of demographic covariates to 0.400916 compared to 0.3347578 in the simple regression model, it is still a relatively low power. If we wanted to have at least a 90% chance of finding a statistically significant difference between the Treatment and Control groups when there exists one at the effect size of 0.3919 more TotalDaysCompleted for the Treatment group, we needed to recruit at least 226 participants, ceteris paribus. For at least 95% power, we needed, at minimum, 284 participants, and for 99% power, we needed 414 participants, all else being equal.

Finally, we conducted tests using the anova function in R to test for differences in group means for all covariates used in both our regression models. The simple linear regression anova results are shown in Figure 28, and complex regression anova results are shown in Figure 29. The results indicate that we fail to reject the null hypothesis that the treatment has no effect on the outcome variable TotalDaysCompleted after controlling for the control group due to very large p-values for all included covariates, providing evidence that we fail to reject the null. In other words, any calculated non-zero difference in group means between the Treatment and Control groups for any included covariates is not statistically significant.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Treatment	1	1.533447	1.533447	0.5422771	0.4649316
Residuals	50	141.389630	2.827793	NA	NA

Figure 28: ANova Results for Reg (Simple)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Treatment	1	1.53344729	1.53344729	0.51091281	0.4784359
Age	1	3.32091846	3.32091846	1.10646111	0.2984684
Male	1	0.03641925	0.03641925	0.01213414	0.9127762
Caucasian	1	2.84970553	2.84970553	0.94946274	0.3350652
Asian	1	0.05141923	0.05141923	0.01713182	0.8964469
Bay	1	0.06873333	0.06873333	0.02290052	0.8803924
Residuals	45	135.06243383	3.00138742	NA	NA

Figure 29: ANova Results for Reg1 (Covariates)

Over the course of the experiment, we tracked the total number of people that completed all 5 days of our assigned tasks in total, and the results are as follows: Day 1 = 52 people, Day 2 = 41 people, Day 3 = 33 people, Day 4 = 30 people, and Day 5 = 24 people. Thus, just under half of all recruited participants completed the entire study. This indicates that most participants were not able to build the 5-day habit of completing the assigned 20 jumping jacks per day. When comparing the Treatment versus the Control groups, out of the 24 people who completed all 5 days, 12 were assigned to Treatment and 12 were assigned to Control. This suggests that there isn't a sizable difference in incentivizing people via daily/immediate payouts compared to lump-sum payments for those subjects who completed all 5 days of tasks in our experiment.

Conclusion

In conclusion, despite power concerns, this study has shown promising signs that engaging people in habit-building behaviors through daily or lump-sum reward systems is possible. Further investigation into various reward structures could lead to the development of versatile applications suitable for numerous scenarios. The potential commercial implications include using such systems to increase customer retention and engagement with businesses, while in the health sector, these findings could help promote healthy habits like regular exercise or improved dietary choices. Moreover, exploring different aspects of the reward system, such as adjusting the size, type, or frequency of rewards, could reveal even more effective strategies for motivating individuals. Although this study has contributed valuable insights into the effectiveness of daily reward systems, further research can help refine these methods and expand their applicability across various domains and contexts. However, due to our experimental results being statistically insignificant, we discuss limitations and potential questions of the study below.

Questions and Limitations

The study has some limitations that are worth noting. Firstly, the experiment was constrained by the $500 incentive available for test participants, which limited the amount of participants we could recruit, the length of the experiment, and potentially could've helped reduce the attrition rate through the test. The reach of our audience was also limited during participant recruitment, making it challenging to create a truly representative sample, which could impact the generalizability of the results. Another concern is that, due to the time constraints of the academic semester, we were not able to examine a test period of more than a few consecutive days, which may undermine the habit-building behavior that is commonly associated over a 21 day time period. Lastly, because we required participants to submit video evidence as the sole means of verifying daily task completion, some potential candidates chose not to participate in the study for a myriad of reasons, including privacy concerns and simply not wanting to leave a digital trail of the task completion to be used by the researchers. This reliance on video proof might have introduced bias by deterring certain individuals who either couldn't or wouldn't provide such evidence despite their interest in the experiment. Increased time available for the study and monetary resources could help mitigate these limitations in future experiments.

Several aspects could be explored in future studies to improve the understanding of reward systems and their effectiveness on task completion. Firstly, it would be valuable to investigate alternative methods of incentivizing users beyond daily or lump-sum rewards, as other systems might prove more effective in motivating participants and generating higher Average Treatment Effects. Secondly, conducting a longer test would allow for habits to develop more robustly, potentially offering more accurate insights into the influence of reward structures on task completion rates. Recruiting a wider participant pool from more diverse backgrounds compared to the ones recruited from the researchers' social circles will improve the representation of folks in the experiment and increase generalizability to the population at large. Lastly, given the constraints of this test, employing a different daily task may have increased the number of potential participants and consequently resulted in a study with higher power, which could contribute to more reliable and generalizable findings.

Various ethical considerations were taken into account to ensure the well-being and privacy of participants. Firstly, the video evidence collected for confirmation of task completion was neither stored nor viewed beyond the initial verification process. Additionally, when using example videos in this report, explicit consent was obtained from the participants involved. Although the monetary incentive could potentially be seen as exploitative in some cases, given the size of the reward and the participant sample, we believe that an exploitative environment was not created via informed consent by the participants. Lastly, recognizing that the chosen task of doing jumping jacks may not be accessible for everyone, future studies may want to consider alternative activities that can be performed by a wider range of individuals to ensure greater inclusivity and accessibility. Overall, while our experimental study faced challenges, noteworthy results were generated and the field study was a valuable learning experience for the team.

Bibliography

Cheney CD, Pierce WD. Behavior analysis and learning. 5th. New York, NY: Psychology Press; 2013. [Google Scholar]

Frömer, R., Lin, H., Dean Wolf, C.K. *et al.* Expectations of reward and efficacy guide cognitive control allocation. *Nat Commun* 12, 1030 (2021). [Nature.com]

"Gym Membership." *Google Trends*, Google, trends.google.com/trends/. Accessed 22 Apr. 2024. [Google Trends]

Matthew Rand a, et al. "Why Do New Members Stop Attending Health and Fitness Venues? The Importance of Developing Frequent and Stable Attendance Behaviour." *Psychology of Sport and Exercise*, Elsevier, 6 Aug. 2020. [ScienceDirect]

Newman, Andrew H., Tafkov, Ivo, Waddoups, Nathan, Xiong, Grazia, The Effect of Reward Frequency on Performance under Cash Rewards and Tangible Rewards (March 30, 2022). [SSRN]

Silverman K, Jarvis BP, Jessel J, Lopez AA. Incentives and Motivation. Transl Issues Psychol Sci. 2016 Jun; 2(2): 97–100. [National Library of Medicine]