Stepping Into The Crucible: How Participants React to Positive vs Negative Treatment During a Walking Competition

Lisa Barceló, Giles Ochs, Hyera Moon

Introduction

Almost 300 million people in the United States alone own a smartphone[1]. And an estimated 40 million people[2] own smartwatches or fitness bands. There are dozens of step tracking apps available in the Apple and Android app stores, and still others available on lesser known platforms like the Windows phone. The majority of smartphones come equipped with default step tracking apps; in the iPhone, this is the Health app. The ubiquity of wearable technology has made it easier than ever to keep track of health and fitness metrics such as steps taken, hours spent standing, and hours spent sleeping. Increasingly, studies regarding user engagement and causal behavior modification have explored the factors related to healthy behaviors, as well.

Getting people motivated to modify their lifestyle remains a difficult endeavor. Much has been written on the efficacy of positive versus negative reinforcement as it pertains to the effectiveness and duration of a lifestyle modification. Human nature, conveniently, prefers to believe that we are most responsive to positive feedback, especially short-term rewards. It is much more pleasant to motivate oneself to go for a long run if you know a spaghetti feed will follow. Less pleasant is the threat of having your roommate cash the $100 check you left as a "motivation" in case you did not follow through. A recent MacArthur Genius Grant application brought together a slew of experts to engineer a large-scale, long-term study (with a budget of $10 million) that would use tangible reward systems to motivate health behavior change. The lead researchers on the team felt if individuals perceived a net positive change in their life as a result of the behavior modification (i.e., going to the gym more often, purchasing healthier foods), they would be more likely to incorporate the change into their long-term lifestyle[3]. Ultimately, this application was unsuccessful, and we are left to our own suppositions as to the potential outcome.

Attempting to motivate behavior change through the use of negative feedback is often seen as the tough, but sure-fire solution to solve the problem of complacency and underperformance. Referring back to our earlier example of how to motivate behavior change, when comparing the positive reward of a spaghetti feed and the negative consequence of losing $100, it seems clear that the specter of financial loss would more than likely be the final push we need to get out the door. In the competitive world of sports, music, and academics, the vast majority of coaches and teachers often resort to shouting, intimidation, and expressions of deep disappointment as a way to drive outcomes. However, excellent outcomes have also been achieved with encouragement and positive rewards. Curious, we decided to see whether or not we could replicate a competitive environment, apply different treatments, and observe outcomes. We hypothesized that we would definitely see a marked increase in the steps for our *negative* treatment group, relative to our control group, and a notable increase for our *positive* treatment group as well.

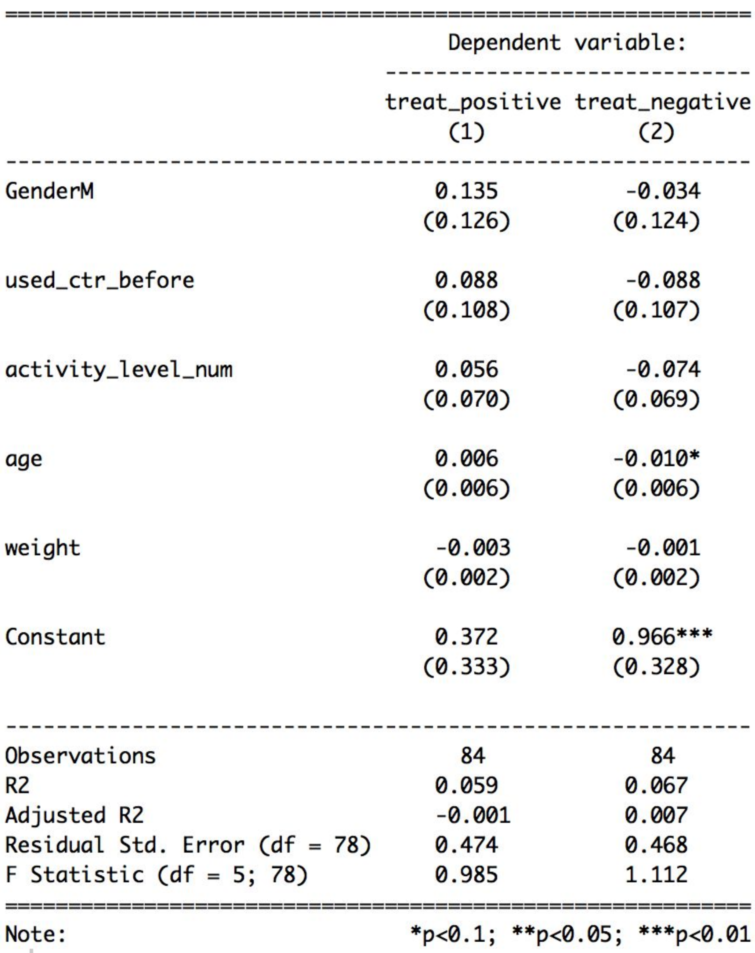
Research Question

Our study sought to explore the effect of positive versus negative reinforcement in the context of a two-week step counting competition. We sought to quantify the impact the various treatment types would have on the trajectory of daily steps taken by individuals in each experimental group during the competition.

Experiment Setup

They say fortune favors the prepared, and we aimed to be as ready as possible. We completed a small, 3-day pilot of our experiment to iron out logistic errors and test several step counting apps, as well as familiarize ourselves with the MailChimp system. During this pilot, we established roles for each member of the research team, and did several dry runs to ensure a smooth flow. Following a successful pilot, we turned our attention to the most difficult part yet: recruitment.

We recruited 90 participants from our academic community and social networks with no respect to age, gender, or activity level. Participants were excluded from the study if they did not have a compatible iPhone device, able to download the required app. Our recruits were randomized into 3 groups, and were asked to complete a demographic survey, which would ultimately yield our pool of covariates. To minimize spillover, married couples, coworkers, and roommates were clustered by home address or workplace. Just before the treatment began, the 7-day average was collected from each participant prior to the start of the experiment as a baseline metric. Our covariates passed the balance test (**Table 1**), and the distribution was even throughout the treatment groups (**Figure 1**).



```
===========================================================
                             Dependent variable:
                        -----------------------------
                        treat_positive treat_negative
                             (1)            (2)
-----------------------------------------------------------
GenderM                     0.135          -0.034
                           (0.126)        (0.124)

used_ctr_before             0.088          -0.088
                           (0.108)        (0.107)

activity_level_num          0.056          -0.074
                           (0.070)        (0.069)

age                         0.006          -0.010*
                           (0.006)        (0.006)

weight                      -0.003         -0.001
                           (0.002)        (0.002)

Constant                    0.372          0.966***
                           (0.333)        (0.328)

-----------------------------------------------------------
Observations                 84              84
R2                          0.059          0.067
Adjusted R2                 -0.001         0.007
Residual Std. Error (df = 78) 0.474        0.468
F Statistic (df = 5; 78)    0.985          1.112
===========================================================
Note:                        *p<0.1; **p<0.05; ***p<0.01
```

Table 1

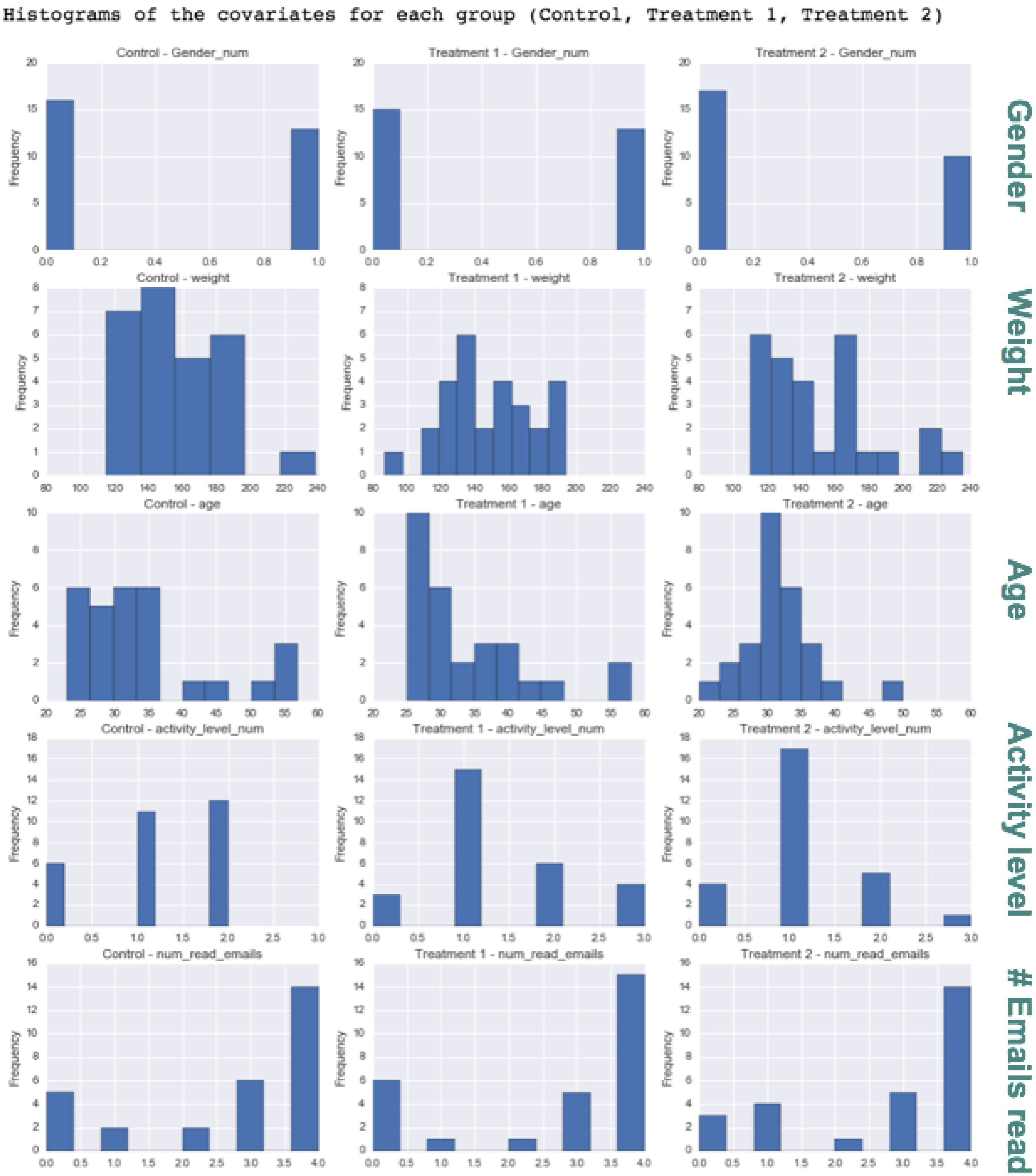


Figure 1

Individuals were told that their daily step count would be collected from the Stepz app (**Figure 2**), and they would be receiving status updates every two days alerting them of their rank. During the course of the competition, participants could not see the rankings of their fellow competitors on the app, only their own.

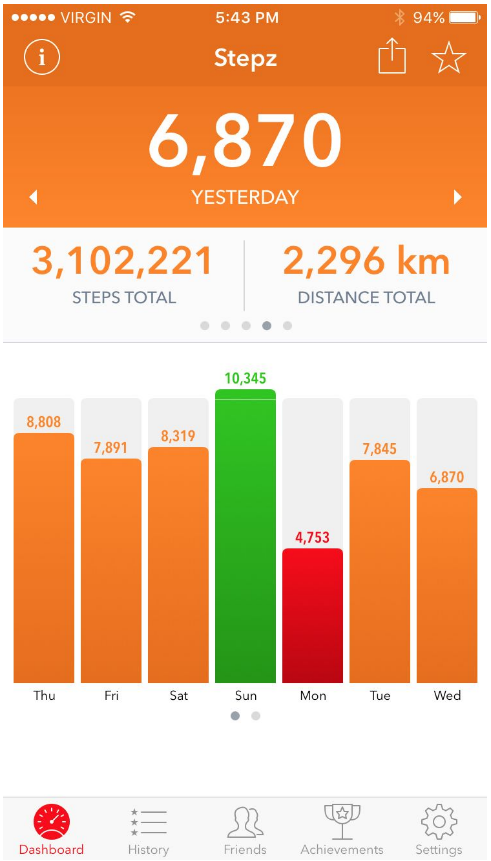
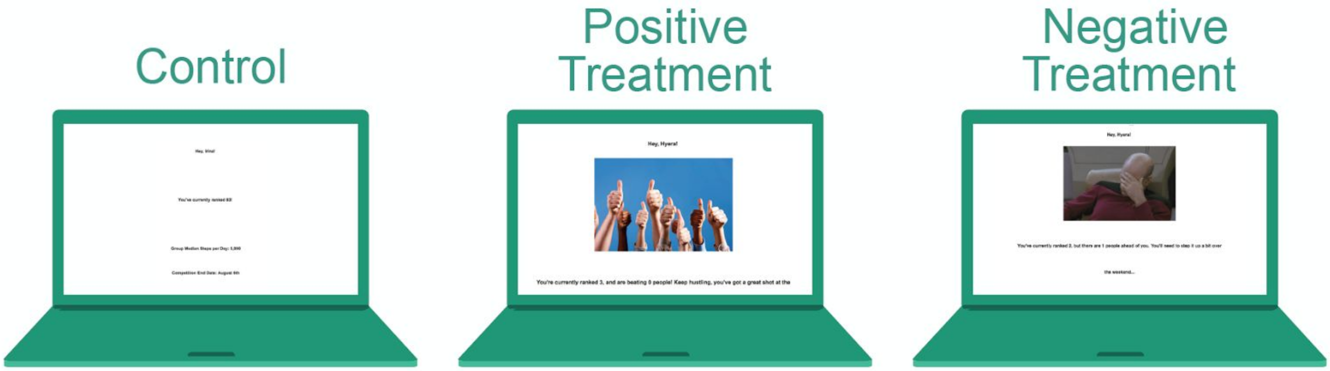


Figure 2



Hey, Vinal

You've currently ranked 83

Group Median Steps per Day: 5,000

Competition End Date: August 6th

Hey, Hyera!

You're currently ranked 3, and are beating 8 people! Keep hustling, you've got a great shot at the

Hey, Hyera!

You've currently ranked 2, but there are 1 people ahead of you. You'll need to step it up a bit over

the weekend...

Figure 3

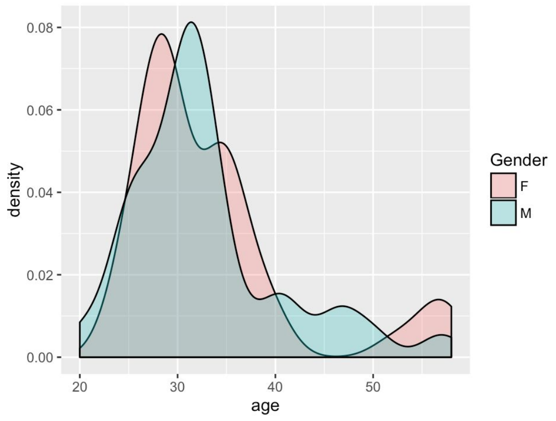
All emails were sent out at the same time of day (**Figure 3**). Our initial "welcome" email was sent out the first day of the competition, and was not considered a treatment. The *control* group received an email with their rank, and how many days remained in the competition. The *positive* treatment group included a positive image (i.e., a picture of a "thumbs up"), as well as an encouraging statement regarding their ranking. Regardless of where the participant ranked, they were told "You are ranked *x* [out of *n*], you are ahead of [*n*-*x*] people!" The *negative* treatment group was given an image showing disappointment (i.e., a "facepalm"), and a message phrased in terms of how many people were ahead of them, i.e., "You are ranked *x* [out of *n*], but there are still [*x* - 1] people ahead of you." It is worth noting that throughout the course of the competition, we never had a last-place candidate in the positive treatment group. Similarly, by random chance, there was never an individual in the negative treatment group that came in first place.

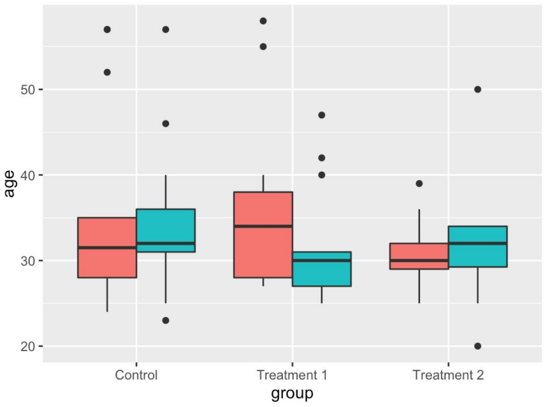
The emails to each treatment group were sent as an "email campaign" that could be tracked using MailChimp. Compliance was measured by whether or not the participants opened the emails. Attrition was measured by the number of participants who either shut off their app, or left their phones at home for a given day, as evidenced by a result of "0" steps for a given day. We found no relationship between attrition and treatment group.

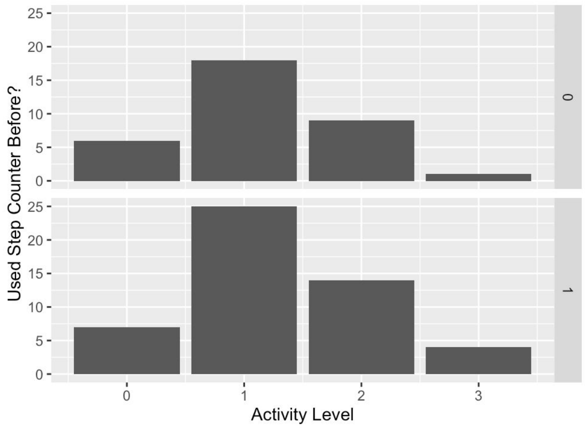
Data Description

1. Demographic data

Our demographic data was collected through a Google survey administered to participants prior to enrollment in our study. We inquired after gender (M/F), weight (lbs), activity level (low = 0, medium = 1, high = 2, really high = 3), and whether or not the individual had used a step counter before (Y=1, N=0). The study population skewed young (Figure 1A), but there was good age and gender distribution throughout our treatment groups.







2. Steps data

Our steps data were collected through the Stepz app[4]. Participants were given instructions on how to create usernames and add our study "W241" as a friend. Regrettably, we did not find a free app that would allow us to mass import steps from our competition participants, so we opted for the vintage approach. We would screenshot the daily steps and log them in our tracking sheet. At the end of the study, to troubleshooot any unforeseen app glitches, we asked participants to send us the final screenshots of their entire two week period so we could do a final reconciliation. It was a worthy exercise. Though only a couple of errors needed remedying, one of these fixes ended up launching an applicant from the bottom of the rank to the top 10 winners!

One main difference we noticed between the participants in the treatment and control groups was their baseline average steps count (**Figure 4**). Participants in the positive feedback treatment group had a higher average steps than the ones in the control group, even before the step challenge first treatment (see Figure 4 on the left side of July 28 dotted line). In the opposite, participants in the negative feedback treatment group had a lower average steps count than the ones in the control group before the challenge. It is to be noted that the treatment was randomly assigned and we did not know which participants had higher or lower step counts. Thus, this observation of participants in positive treatment with higher pre-challenge steps and participants in negative treatment with lower pre-challenge steps resulted from a random process and is not reflecting any effect from the positive and negative reinforcement causal effect we are interested to study in the experiment.

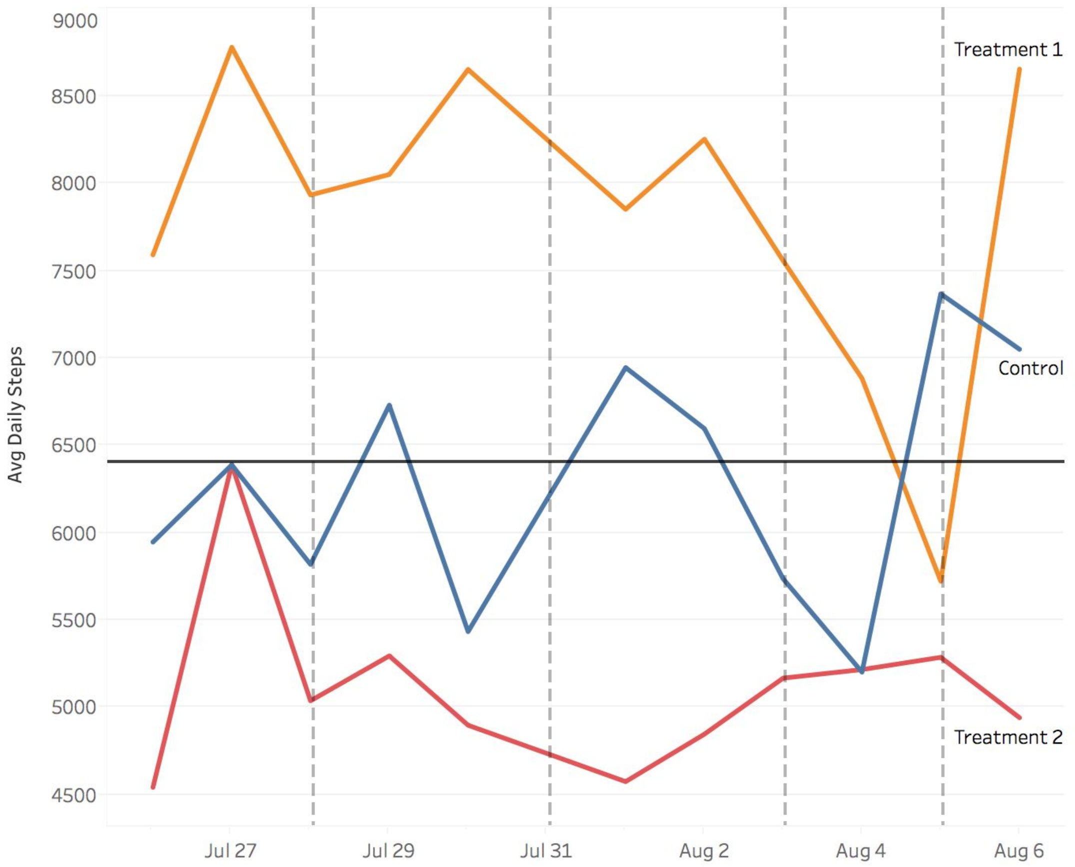


Figure 4: Average daily steps per control and treatment groups

Note: The first dotted vertical line (July 28), the first day the positive and negative treatment and placebo emails were sent to participants.

Methods and Findings

We are interested in measuring the change in number of steps from the positive or negative treatment emails, that is, a difference in differences, as shown in Figure 5 and Figure 6 below.

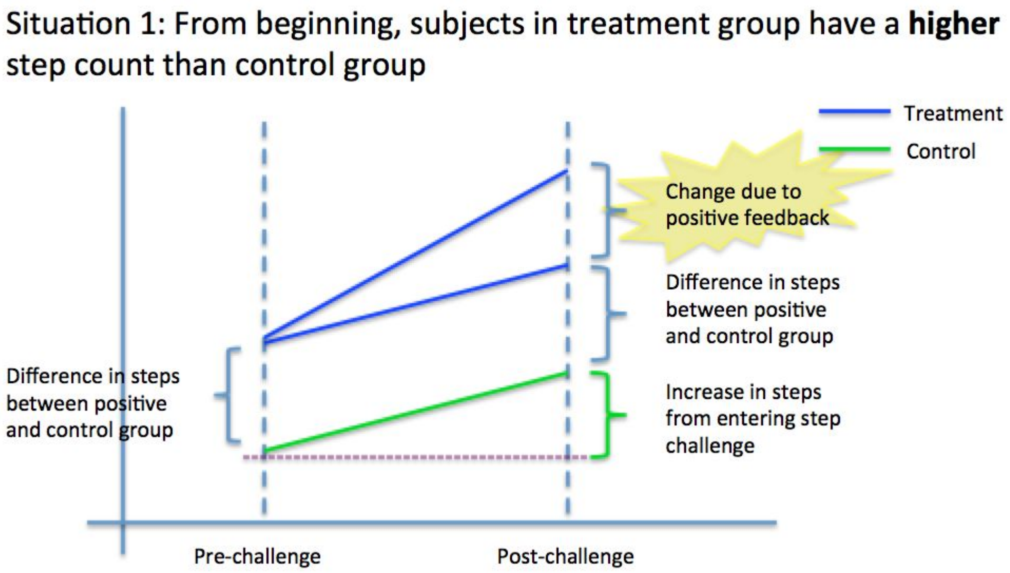

Situation 1: From beginning, subjects in treatment group have a **higher** step count than control group
Treatment
Control
Change due to positive feedback
Difference in steps between positive and control group
Increase in steps from entering step challenge
Difference in steps between positive and control group
Pre-challenge
Post-challenge

Figure 5

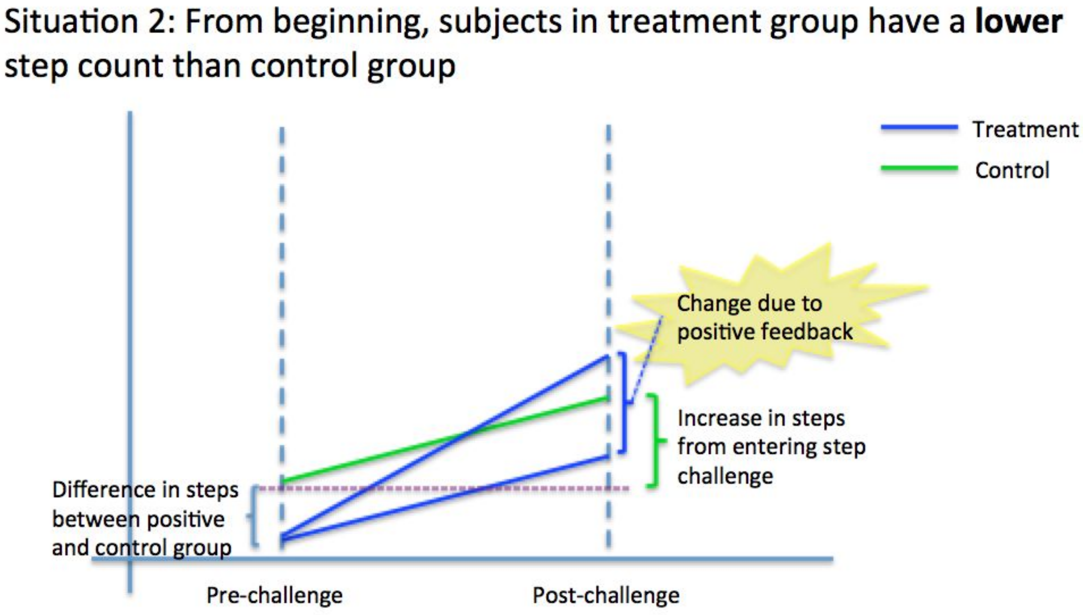

Situation 2: From beginning, subjects in treatment group have a **lower** step count than control group
Treatment
Control
Change due to positive feedback
Increase in steps from entering step challenge
Difference in steps between positive and control group
Pre-challenge
Post-challenge

Figure 6

Based on a difference in difference model, our linear regression has been developed as follows:

```
step_avg_prepost  ~ 1
```

```
                    + treat_positive + treat_negative
                    + time_post
                    + treat_positive*time_post
                    + treat_negative*time_post
```

where

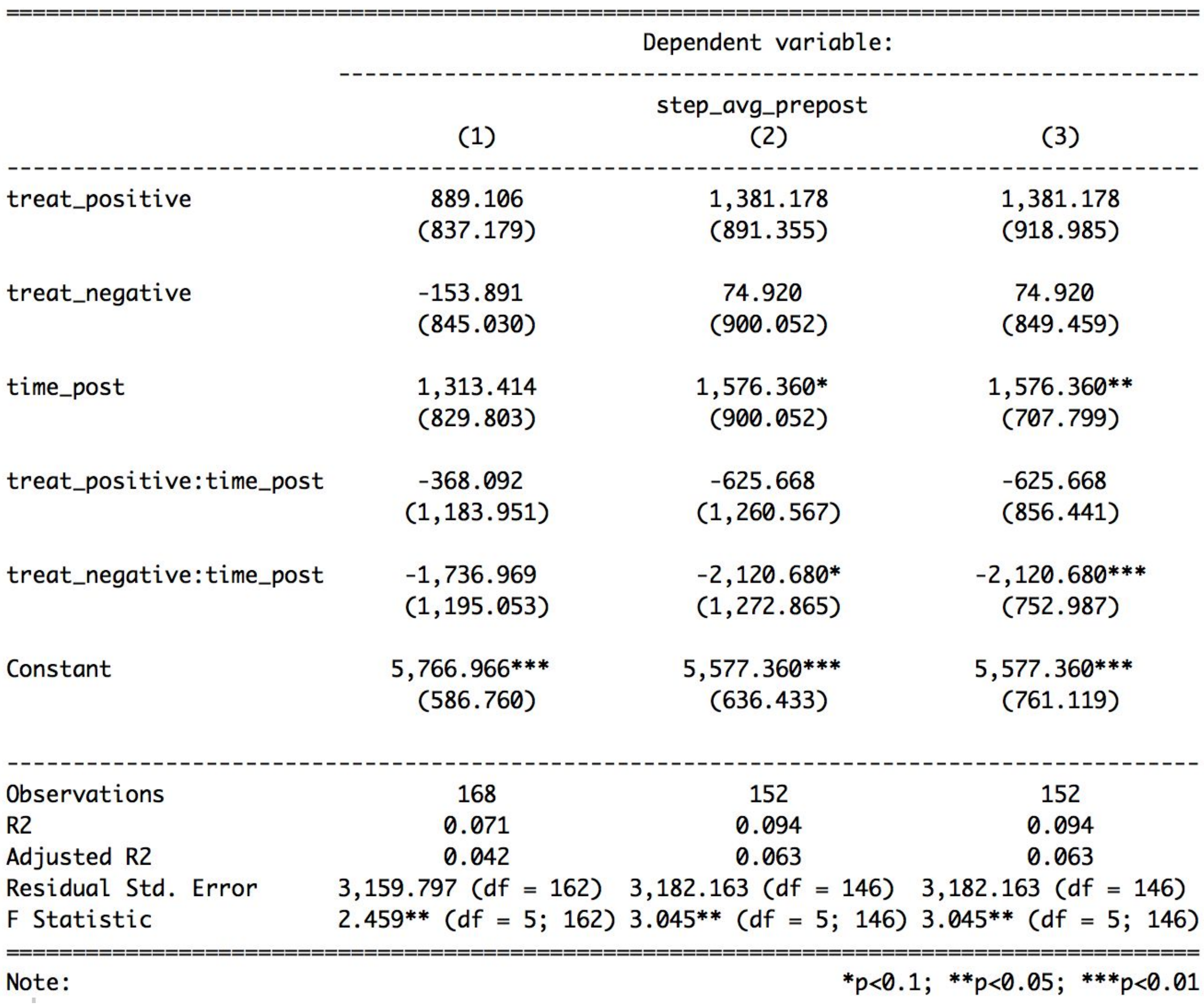
- `step_avg_prepost`: 7-day pre-challenge period average steps;
- `treat_positive`: **1** if not in treatment group for positive feedback email and **0** otherwise
- `treat_negative`: **1** if not in treatment group for negative feedback email and **0** otherwise
- `time_post`: **1** if observation from challenge period, **0** if from pre-challenge period

The estimate of the effects shown in **Figure 5** and **6** will be represented by the coefficients in the linear regression:

- Increase in steps from entering the step challenge: coefficient of variable `time_post`
- Difference in steps between positive and control group: coefficient of variable `treat_positive`
- Difference in steps between negative and control group: coefficient of variable `treat_negative`
- Change due to positive feedback: coefficient of the interaction term between `treat_positive` and `time_post`
- Change due to negative feedback: coefficient of the interaction term between `treat_negative` and `time_post`

We built this model with 3 variations as shown in **Table 2** below.

- **model_1**: We started with all 84 subjects individual data points for each pre-challenge and step challenge periods.
- **model_2**: There were 8 non-compliers in the experiment; that is, 8 subjects did not read any emails (placebo or treatment emails) during the entire step challenge period. Since our control group were also given a placebo email, we were able to identify the non compliers for both control (placebo) and treatment groups. Thus, we removed the non-compliers from the analysis to study the complier average causal effect (CACE).
- **model_3**: In addition to the CACE from **model_2**, to reflect the clustering of individual subjects when randomly assigning placebo and treatment groups, we used clustered standard errors.



```
==============================================================================================
                                            Dependent variable:
                         ---------------------------------------------------------------------
                                              step_avg_prepost
                                 (1)                    (2)                    (3)
----------------------------------------------------------------------------------------------
treat_positive                 889.106              1,381.178              1,381.178
                              (837.179)              (891.355)              (918.985)

treat_negative                -153.891                74.920                 74.920
                              (845.030)              (900.052)              (849.459)

time_post                     1,313.414             1,576.360*             1,576.360**
                              (829.803)              (900.052)              (707.799)

treat_positive:time_post      -368.092               -625.668               -625.668
                             (1,183.951)            (1,260.567)             (856.441)

treat_negative:time_post     -1,736.969            -2,120.680*            -2,120.680***
                             (1,195.053)            (1,272.865)             (752.987)

Constant                     5,766.966***           5,577.360***           5,577.360***
                              (586.760)              (636.433)              (761.119)

----------------------------------------------------------------------------------------------
Observations                    168                    152                    152
R2                             0.071                  0.094                  0.094
Adjusted R2                    0.042                  0.063                  0.063
Residual Std. Error      3,159.797 (df = 162)  3,182.163 (df = 146)  3,182.163 (df = 146)
F Statistic             2.459** (df = 5; 162) 3.045** (df = 5; 146) 3.045** (df = 5; 146)
==============================================================================================
Note:                                                        *p<0.1; **p<0.05; ***p<0.01
```

Table 2

Our results suggest that there was a statistically significant effect from the negative treatment group. The size of the estimated is important: participants in the negative treatment group showed an overall decrease of 2,121 steps per day, about a 40% drop from the pre-challenge average steps of 5,577 steps. This effect runs counter to our initial hypothesis of negative reinforcement leading to an increased step count.

Also, an interesting result from the experiment is that the coefficient of the variable `time_post` turns out to be statistically significant. This coefficient represents the change (+1,576) in average daily steps between the pre-challenge and challenge period, independent of the treatment effect. This suggests that merely participating in the competition itself led participants to be more active than they normally would.

Taken together, our findings indicate that there is a benefit to being a part of a competitive group wellness activity. However, our analysis indicates that a negative reinforcement does not drive better outcomes - in this case, an increase in number of steps - but results in lower performance.

Discussion and Future Directions

In spite of our prep work, pilot study, and interesting findings, there are still a few areas of concern around recruitment, treatment methods, and attrition.

We recruited participants primarily from among our friends and colleagues, limiting the generalizability of the study. People tend to associate with other people from similar backgrounds, and thus we cannot make sweeping statements about how all, or even most, people can be expected to perform under pressure.

Our treatment vehicle, email, is an imperfect proxy for the consequence and reward systems that abound in the real world. Regardless of which type of incentive is being applied (i.e., receiving a promotion at work or being yelled at by a coach on the ball field), the in-person effect is significantly more impactful than an email. In order to mitigate the muted effect our electronic messages, our team initially considered strengthening the language of our treatment emails, specifically the negative treatment. In the end, we felt it would be unethical to prey on the willingness of our colleagues and friends and subject them to anything harsher than Captain Picard's emphatic facepalm of disappointment.

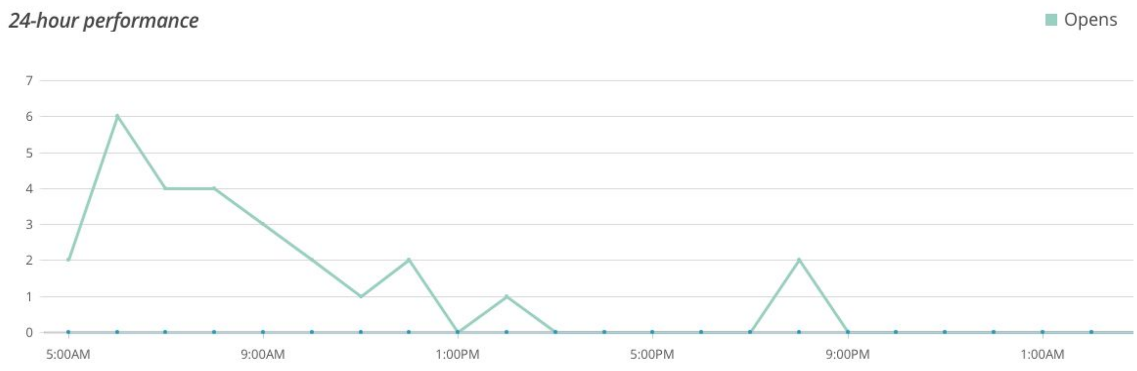
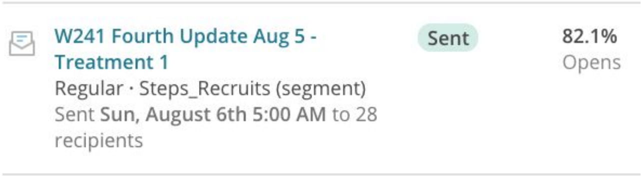


Figure 7. MailChimp shows a timeline of our opens so we can see the peak engagement.


W241 Fourth Update Aug 5 - Treatment 1
Regular · Steps_Recruits (segment)
Sent Sun, August 6th 5:00 AM to 28 recipients

Sent

82.1%
Opens

Figure 8. For each email sent out, we can see how many people have opened it.

A second concern surrounding email as treatment method is that, while MailChimp allows us considerable insight into the activity of the individuals on our mailing list (**Figure 7**), the truest measure of treatment still eludes us. As was pointed out by a couple of our participants, receiving emails on iPhones with smaller screens means when you open the email, you are only exposed to the top portion! One individual (in the control group) mentioned to us that she didn't ever scroll down to see the rest of the message, only her ranking. It then occurred to us that perhaps there was differential exposure to our treatment, even among the compliers, although we were ill equipped to study the differences.

Hindsight is often 20/20, and ours is no exception. Had we more time, an endless budget, and access to a plethora of sports equipment, we would love to replicate our study in a truly competitive environment, allowing for a stronger dosage of each treatment to be applied. It goes without saying that our current method of measurement (screenshotting steps and entering them into a spreadsheet) was valiant, but not scalable in the slightest. Ideally, we would use a piece of wearable tech delivering data to us through a live feed.

References

1. http://www.pewinternet.org/fact-sheet/mobile/
2. http://www.mobihealthnews.com/content/study-12-percent-us-consumers-own-fitness-band-or-smartwatch
3. http://freakonomics.com/podcast/solving-one-problem-solve-others/
4. https://itunes.apple.com/us/app/stepz-pedometer-step-counter-for-tracking-steps/id839671656?mt=8