# Effect of Fitness Competitions on Activity Levels

## Spring 2019 - W241 Final Report

[redacted]

e

## Abstract

Employee wellness programs offered by corporations intend to incentivize and motivate employees to take preventative actions to improve their health and reduce medical spending. Many people in the United States have access to some variation of a wellness program and companies collectively invest billions of dollars annually in hopes of achieving both desirable outcomes. Do employee wellness programs actually cause improved health and reduced medical spending? In this project, we study the impact that a wellness program may have on people.

We do this by recruiting colleagues, friends, and family to share their daily steps counts with us. We intervened with the treatment group during the second week of the study in order to perform a difference in differences assessment. The intervention was a step competition, which helped us measure the additional steps that individuals take if they are in a competition with people they know. In the results, we found there to be an increase in step counts for people in a competition. nice setup!

## Background

meh...

Corporations offer Employee wellness programs with the intention to incentivize and motivate employees to take preventative actions to improve their health and reduce medical spending. The definition of an employee wellness program from HealthCare.gov can include components like health screenings, reimbursement for gym memberships, and fitness initiatives. Many people in the United States have access to some variation of a wellness program and companies collectively invest billions of dollars annually in hopes of achieving both desirable outcomes. So is the investment warranted? Do employee wellness programs actually cause improved health and reduced medical spending? Are particular implementations more effective at causing these two outcomes? Both previous observational studies and randomized control studies attempt to answer some aspects of these broad questions. In this project, we want to study the impact that a more focused element of wellness programs may have on people.

If you're going to include this ¶, then you should lean into it more. Include cites, or @ least high-level takeaways.

## Research Question

During a 14 week course and a single field experiment, the broad questions of understanding the impact of wellness programs is too ambitious to definitively answer. However, a particular implementation of one component can be studied. Some companies promote fitness competitions as the trademark event in their fitness initiatives. When offering fitness competitions, companies seek to create excitement and community around an activity, like walking and running, which can be measured in a step competition. Step competitions are appealing because increased steps can be incorporated into a daily routine while also being consistently measurable.

ok

The scope of this research project is to broadly study individuals and not just people who are part of a single company. Additionally, we are studying the changes in people's step count habits, and not specifically measuring improvements in health or medical costs.

### Hypothesis

The question we attempt to answer in this experiment is: Does competitive step tracking increase weekly individual step counts?

→ & Do you have a theory for why?

We ~~optimistically~~ predict that the monetary incentives and social interactions provided by the leaderboard challenge will increase the number of steps taken. We believe that over a brief 5 day period, we can set up the prize money scheme, competition size, and dynamics to measure this effect.

## Experiment Design

### Experiment Overview

To test this hypothesis and address the research question, we design a difference-in-differences experiment. All participants have their steps measured throughout the 2 week study period, then random assignment occurred on the Saturday after week 1 where the competition intervention is delivered to the treatment group. The core analysis compares the difference in weekday daily step counts in week 1 and week 2 for the treatment and control groups.  Below is the exact project timeline. We also lay out the measurement tooling, communication tooling, enrollment and recruitment process, randomization, and observation and outcomes.

## Project Timeline

| **Start Recruitment** | **Send out App Sign Up** | **Week 1 Baseline Data Collection** | **Randomization and Treatment Delivery** | **Week 2 Baseline Data Collection** | **Participant Debrief & Awards** |
|---|---|---|---|---|---|
| Friday March 1 | Wednesday March 13 | Monday March 18 to Friday March 22 | Saturday March 23 | Monday March 25 to Friday March 29 | Sunday April 21 |

## Measurement Tooling - Stridekick

Stridekick is a 3rd party fitness app that allows users to share their steps with a network of friends. Stridekick can connect to a user's device to collect the steps measured on the particular step tracker. This allows for compatibility across platforms, including Garmin, FitBit, AppleWatch, GoogleFit, Apple Health, and others. Stridekick also assists in administering step competitions. Challenge creators can invite friends and set the rules and then Stridekick provides a leaderboard with all participant's most recent cumulative step count.

For this study, we created an account called @MIDS-Steps to administer the treatment challenges and collect data. We chose not to connect a device so that our account always shows "0 steps" for the day.  To see people's steps, we asked them to send us a friend request to connect with our account. On Stridekick, you can only see your own steps and your friend's steps. So if someone is not part of a challenge, then they cannot see other participant's step counts. We could use the app or web portal to create the challenges with 6 to 7 people for the treatment group. Then those people assigned to treatment who accepted invitations can see their fellow competitors.

## Communication Tooling - Qualtrics

We leveraged Qualtrics to contact our participants and gather survey input from them. This platform allows us to maintain email distribution lists, set up reminders, thank you messages, and track responses. The main survey was the Stridekick sign up survey. The gathered information allows us to match email addresses to Stridekick usernames. The survey is also compatible with both desktop and mobile web browser, reducing the barrier to entry for completion.

In the initial sign up, we only collected volunteer names and email addresses. This became problematic when we realized that our messages were getting blocked by some spam filters. In a future study, we would recommend collecting an alternative means of contact like another e-mail address or a phone number. We would also recommend alerting people at the first phase of sign up to look for emails from a particular email address and domain.

→ And, to confirm receipt

## Enrollment & Recruitment Process

The motivation for participants to join our study was two-folder: 1) the potential to win a $25 gift card, and 2) to help out a friend by doing something fun. The treatment group was more incentivized (during the second week) as part of their competition, where first place got a $35 gift card and second place got a $15 gift card. Our first step was to collect signups using a very simple form and a consistent message that each of us used to send over slack, email, texts, etc. Once we had over 100 people signed up, we sent out an email, with an additional reminder, to our participants with instructions on getting setup with the experiment using Stridekick. As a result of our recruiting strategy, our participants turned out to be co-workers, family, friends, and MIDS classmates.

→ OK - so, can you separate the cash motivation from the competition motivati

A two stage design was used for recruitment and enrollment. In the first stage, we sent a consistent message to all potential participants [Appendix A - #1] providing background information and a link to sign up. The Google Form [Appendix B - #1] collects the participant's name and e-mail address along with any other comments or questions. The recruitment message also noted that all participants will be entered into a raffle for $25 Visa gift cards as a sign of appreciation for their time and effort. Our recruitment channels included our work, gym, family, friends, classmates, and other networks including fellow MIDS students. We found that direct messages via e-mail, text message, Slack, and Facebook messenger led to more responses.

Upon collecting these names and email addresses, we used Qualtrics to send a follow up message with the steps to set up Stridekick for the experiment. This also included a detailed set up guide with photos and a Frequently Asked Questions document based on the questions and comments gathered in the Stage 1 Google Form and any other interactions with volunteers [Appendix C]. In addition to configuring Stridekick, we asked that volunteers complete a survey providing their username.

Good

Then throughout the 2 weeks, we sent emails via qualtrics every 2 to 3 days reminding all participants to continue syncing their steps. On Saturday, March 23, the Qualtrics email differed for the control and treatment groups, where the exact messages can be seen in Appendix A sections 3 and 4. In summary, the Qualtrics email to the control group was consistent with the other reminders throughout the data collection period. The treatment group was informed that they will be invited to a competition with an opportunity to win 1 of 2 Visa gift cards for a top performance. We used the surveys in those emails to collect any questions or concerns from participants so they could be addressed quickly.

## Randomization

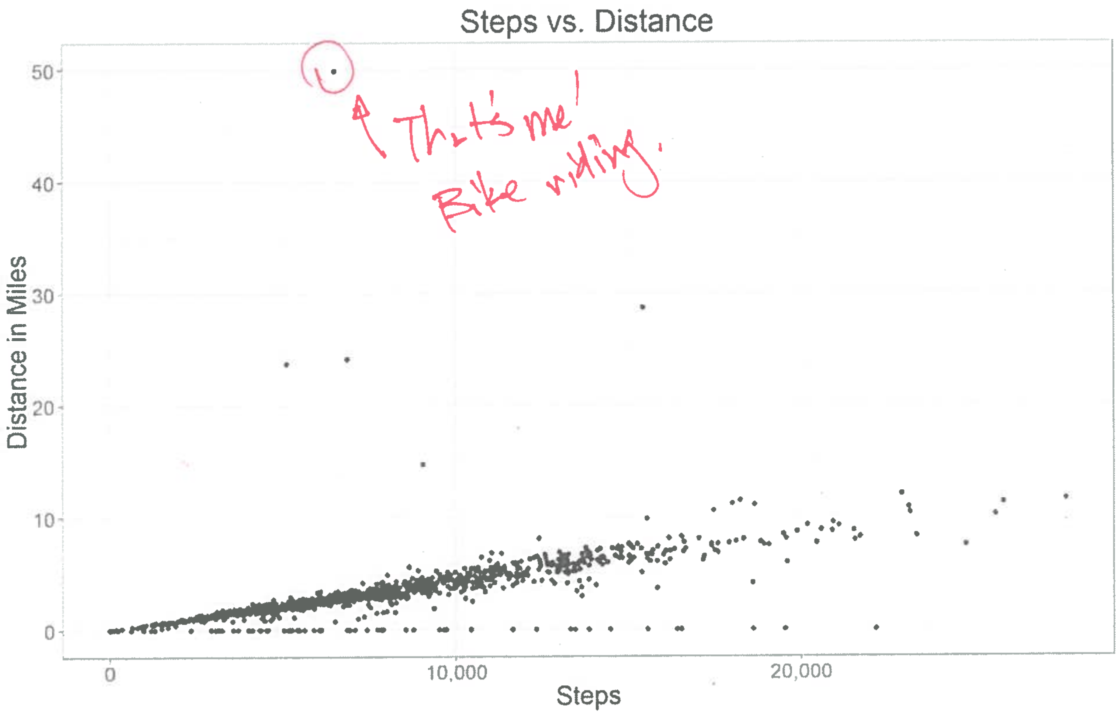
We decided on a clustered randomization approach for two reasons: 1) to avoid spillover since everyone recruited were people we knew; and 2) this works nicely for our intervention of a competition with people you know. People who worked together, were in the same family or friend group, or other relationships were put into the same cluster. There was one large cluster and many clusters of size 1-4.

After assigning clusters, we build a sampling randomization function in R to randomly assign each cluster to control or treatment. This was followed by placing each treatment cluster into a competition with 6-7 total subjects per competition.

## Observations and Outcome Measures

The data we collected through Stridekick was broken into daily steps and daily distance. This means for each subject, we have their number of steps and distance for each day of the study. We downloaded data for March 18 to March 29, 2019 so that we can compare Monday through Friday of the first week against the second week. While we had some data for the weekend in between, we decided to ignore this data so that we can have an apples to apples comparison of the weekdays from week 1 against week 2.

Since our focus was on steps, we didn't actually ask the subjects to sync distance data. As a result, the distance data was missing a lot of data. While there was some attrition in the steps measure (discussed more later), it was less than the distance data. We also noticed that steps and distance are highly correlated, as shown in the plot below:



For these reasons, we chose to focus primarily on daily steps as our key outcome measure. This was the intention from the beginning of the study.

## Data Completeness

this is a really impressive recruitment effort!

As seen in the flow diagram below, we started with 118 people that signed up for the initial brief survey [Appendix B - #1]. Of those 118 people, 85 people signed up for the Stridekick app and befriended our admin account. We define noncompliance as completing the initial brief survey and not signing up for the Stridekick app. Therefore, 27.97% of people did not comply.

As discussed above, we assigned the 85 compliers to 39 clusters. Then, we assigned 20 of those clusters to the control group and 19 to the treatment group. In all, there were 33 people in the control group and 52 people in the treatment group.

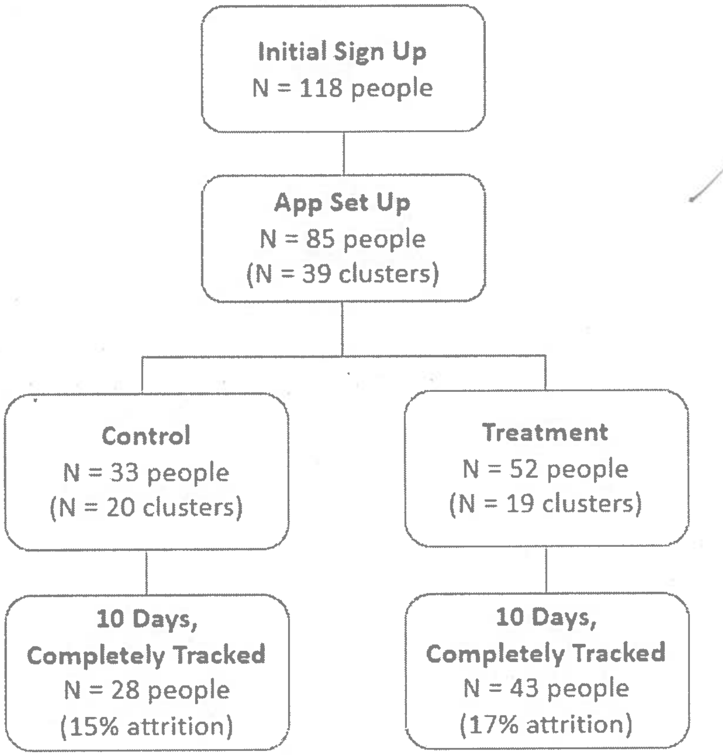
This is pretty good, actually.

We did not receive step counts for every person for every day in `Week 1` and `Week 2`. We are missing 18 observations for 4 unique subjects, with an observation defined as a step count for a given subject on a given day. We have an additional 42 observations that have zero `total.steps` for 12 unique subjects. Given that it's essentially impossible to have zero steps in a day, it is clear these subjects did not "sync" their steps for that day. Therefore, we are missing the essential step count data for these 42 observations.

In total, we are missing 60 observations for 14 unique subjects. Those 14 unique subjects have 108 complete observations. Due to the incompleteness of data for these 14 subjects, we have claimed these subjects attrited. Therefore, we have completely removed all 60 incomplete observations and 108 complete observations for these 14 subjects.

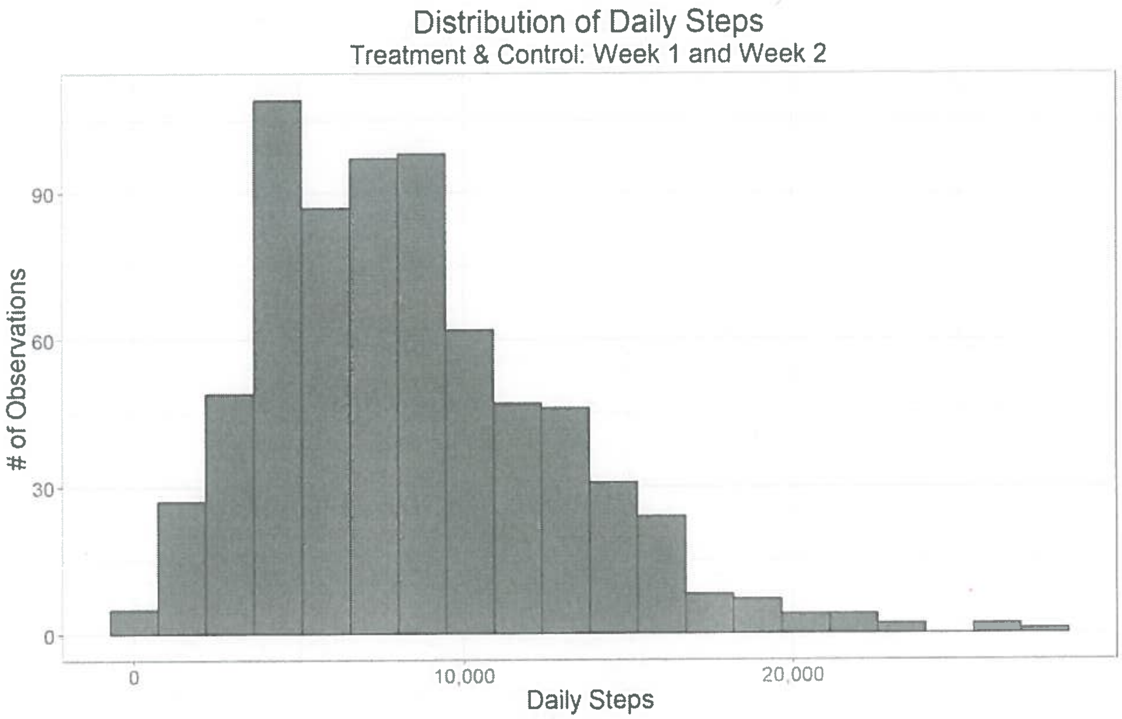
We define attrition as signing up for the Stridekick and not properly recording daily steps for every day in `Week 1` and `Week 2` (i.e., 10 step count observations that are nonzero). After accounting for attrition, our final analysis dataset has 710 daily step count observations, for 71 subjects, for 10 days (`Week 1` and `Week 2`), in 35 unique clusters. There are 28 people in the control group and 43 people in the treatment group. In all, we had 15% attrition in the control group and 17% attrition in the treatment group.

***Treatment v. Control Flow Diagram***



## Results

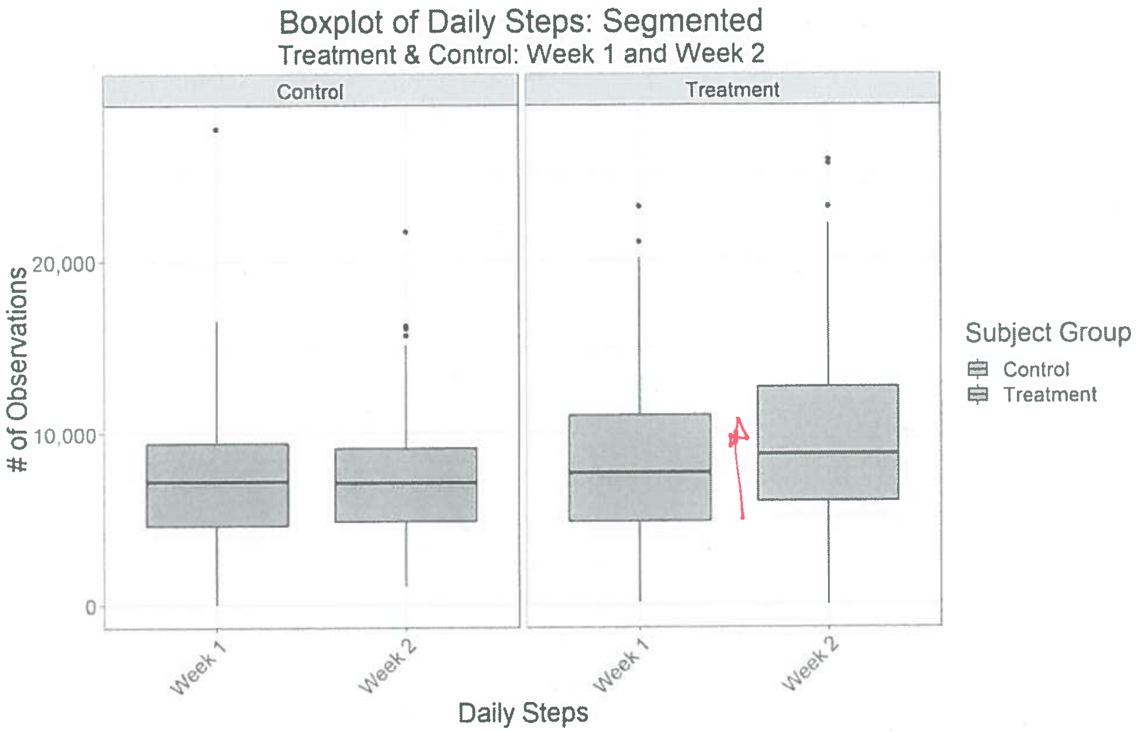
For our analysis, we focus on analyzing the daily steps for each user that has complete data. We analyze the steps for each weekday from `Week 1` and `Week 2`. As seen below, over the course of `Week 1` and `Week 2`, the distribution of steps is right-skewed with a median of 7,766 and a mean of 8,338.



The right-skew distribution is present in every subcategory of treatment/control and week number. It makes sense that everything is right-skewed since some people will have many daily steps and no one can actually have negative steps.

We are mostly concerned with the change in daily step counts from `Week 1` to `Week 2`. As seen below, the median value from `Week 1` to`Week 2` remains relatively constant for the control group, but increases for the treatment group.

Does this lead you toward a data transformation? Perhaps, choosing `log(steps)`?



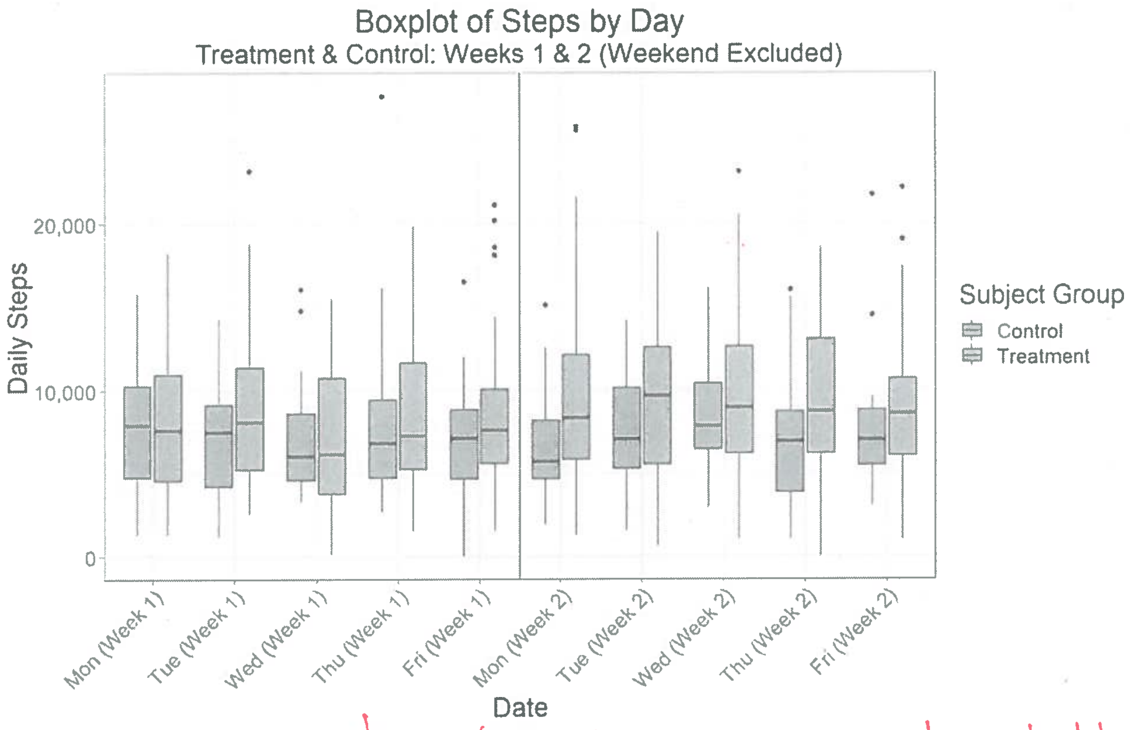
As mentioned earlier, we are focusing our comparisons of steps on `Week 1` versus `Week 2`, ignoring the weekend in between. We can see the from above boxplots that there is some improvement in steps for the treatment group in `Week 2`. However, the treatment group had more steps then control during each week.

As seen in the table below, the percentage difference between the treatment and control groups was larger in `Week 1` (~10.14%) than in `Week 2` (~24.78%). Conversely, the percentage increase in steps from `Week 1` to `Week 2` is much higher for the treatment group (~13.99%), compared to the control group (~0.60%). This suggests that the treatment group increased their average step counts from `Week 1` to `Week 2` more than the control group.

| | Average Daily Steps | | | |
|---|---|---|---|---|
| | **Week 1** | **Week 2** | ***Diff*** | ***% Diff*** |
| **Control** | 7,516.11 | 7,562.44 | 46.32 | 0.62% |
| **Treatment** | 8,278.22 | 9,436.33 | 1,158.10 | 13.99% |
| ***Diff*** | 762.11 | 1,873.89 | | |
| ***% Diff*** | 10.14% | 24.78% | | |

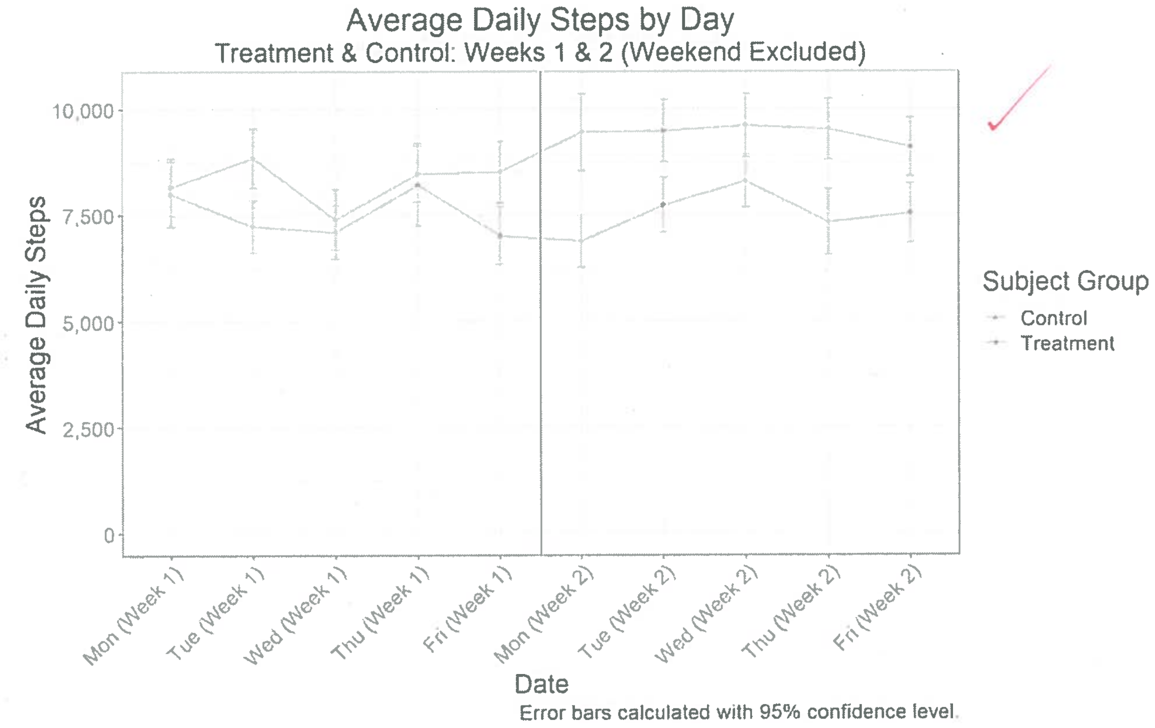
I see what you're shooting for here — but I think this table could be more descriptive. The % change is a little confusing.

As seen below, when breaking down the results by day, one can see the larger deviation between the treatment and control group after the application of the treatment.

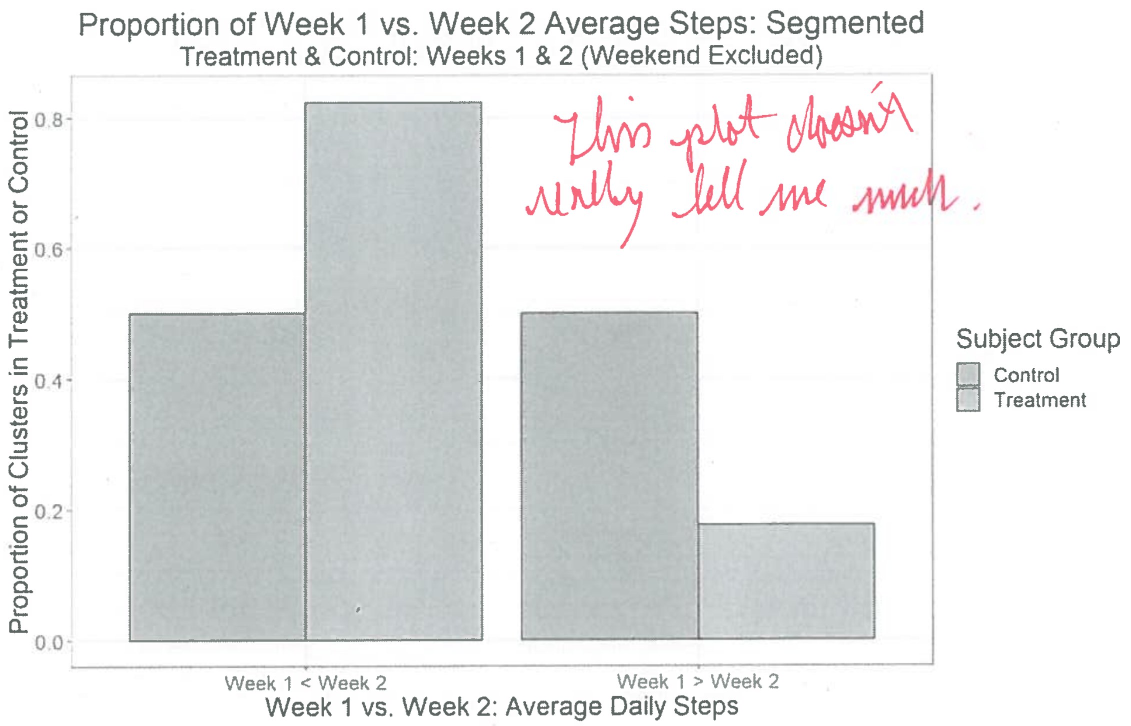


This is AN AWESOME descriptive plot.

This can be seen more directly below when simply viewing the mean steps per day (with 95% confidence error bars) for the treatment and control groups in `Week 1` and `Week 2`. As seen below, while the treatment group outperformed the control group in `Week 1`, the treatment group more drastically and more consistently outperformed the control group in `Week 2`.



The increase from `Week 1` to `Week 2`in the treatment group was not unanimous in every cluster. As can be seen below, half the control group clusters increased their average daily step count from `Week 1` to `Week 2` and half decreased. For the treatment group clusters, over 80% of the clusters increased their average daily step count from `Week 1` to `Week 2` . This suggests that while the treatment had an adverse or beneficial effect, depending on the makeup of the group (i.e., cluster), the majority of clusters had a beneficial effect of treatment.



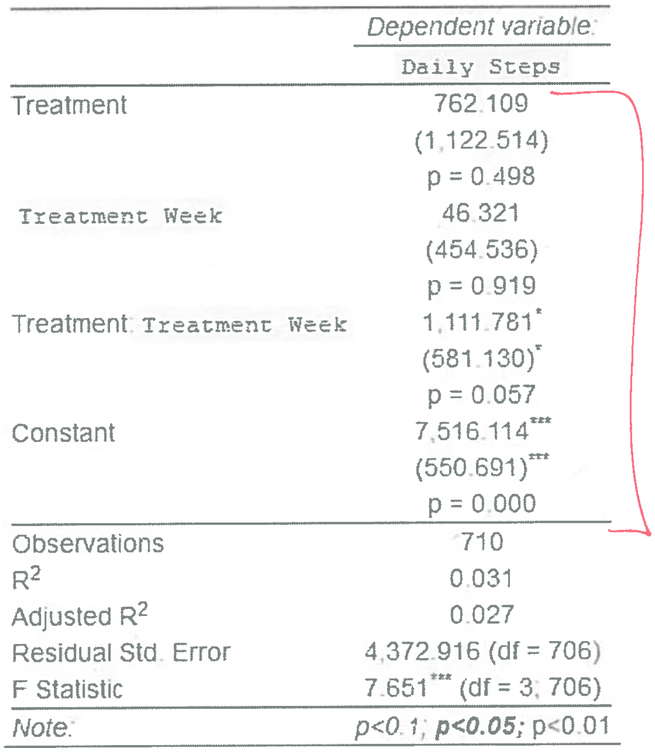
This plot doesn't really tell me much.

Through these visual analyses, we can see that the step competition for the treatment group did have an effect on the subjects doing more steps. We will now move into checking for statistical significance of the treatment effect.

## Regressions

Ultimately, when using a difference-in-differences model on our panel data, we found that the differences-in-differences estimator (interaction term with the treatment application and treatment time period of `Week 2`) shows a statistically and practically significant positive effect of treatment on step counts in `Week 2` for the treatment group. We used a differences-in-differences model because our data was longitudinal in nature and we wanted to estimate the effect of treatment in the treatment and control groups between the time periods of `Week 1` and `Week 2`.

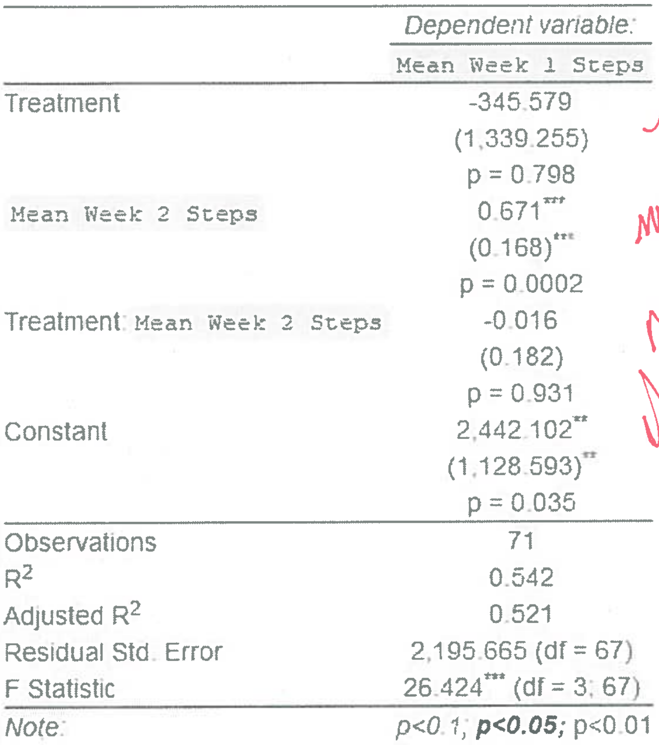
As seen below, the estimated mean difference in daily steps between the treatment and control group, prior to the treatment distribution, is 762.10. This estimated difference is not statistically significant. The estimated mean change in daily steps after the treatment distribution in the control group is 46.32. This estimated change is not statistically significant. The estimated differences-in-differences estimator is 1,111.78 daily steps after the treatment distribution in the treatment group. This estimated change is statistically significant with a p-value of .0561 (calculated with robust cluster standard errors). Meaning, the estimated mean difference in daily steps between the treatment and control group, after the distribution of treatment, is 1,158.10 (46.32 + 1,111.78).



| | *Dependent variable:* |
|---|---|
| | Daily Steps |
| Treatment | 762.109 |
| | (1,122.514) |
| | p = 0.498 |
| Treatment Week | 46.321 |
| | (454.536) |
| | p = 0.919 |
| Treatment: Treatment Week | 1,111.781* |
| | (581.130)* |
| | p = 0.057 |
| Constant | 7,516.114*** |
| | (550.691)*** |
| | p = 0.000 |
| Observations | 710 |
| $R^2$ | 0.031 |
| Adjusted $R^2$ | 0.027 |
| Residual Std. Error | 4,372.916 (df = 706) |
| F Statistic | 7.651*** (df = 3; 706) |
| *Note:* | *p<0.1; **p<0.05;** p<0.01 |

YES YES YES YES!

We also conducted a placebo test by running a linear regression using `Week 1` average steps for a single subject as the dependent variable and `Week 2` average steps for a single subject and `Treatment` as the independent variables. We found that our `Treatment` and the interaction of our `Treatment` with `Mean Week 2 Steps` didn't have statistically significant effects on `Mean Week 1 Steps`. Given that `Week 1` was pre-treatment, we passed this placebo test.



| | *Dependent variable:* |
|---|---|
| | Mean Week 1 Steps |
| Treatment | -345.579 |
| | (1,339.255) |
| | p = 0.798 |
| Mean Week 2 Steps | 0.671$^{***}$ |
| | (0.168)$^{***}$ |
| | p = 0.0002 |
| Treatment:Mean Week 2 Steps | -0.016 |
| | (0.182) |
| | p = 0.931 |
| Constant | 2,442.102$^{**}$ |
| | (1,128.593)$^{**}$ |
| | p = 0.035 |
| Observations | 71 |
| $R^2$ | 0.542 |
| Adjusted $R^2$ | 0.521 |
| Residual Std. Error | 2,195.665 (df = 67) |
| F Statistic | 26.424$^{***}$ (df = 3; 67) |
| *Note:* | *p<0.1;* ***p<0.05;*** p<0.01 |

This is more of a randomization check.

## Conclusions

We can see from this study and the results that we did find a statistically and practically significant impact of the step competition on step counts. Given potential biases and assumptions, our results suggest that our hypothesis is true. Since this study focused solely on step counts as a result of a step competition, we cannot make any assertions about overall health outcomes, as related to the broader research motivation of employee wellness programs' effectiveness. However, this is a solid step forward in better understanding the impact of these programs, and at the very least, we can have confidence in our assessment that these programs result in people being more physically active.

## Limitations and Future Enhancements

There are several limitations of this study that may create some bias or decrease the generalizability of the study's results. Firstly, we only ran this study for two weeks, so it would be a good idea to re-run a similar experiment for a longer period of time. Not only would a longer study give us a better practical confidence in the results, it would also mitigate the novelty problem where people are walking more during the first week than they normally would, just because they're participating in a step tracking experiment. If we could run the study for a really long time, then we can also have better outcome measures directly related to the original research motivation. For example, if we run the study for 6 months, we may consider measuring subjects' change in weight, change in blood sugar, or other health measures.

Another limitation is the sample of subjects. The people recruited for this study were colleagues, friends, and family of 3 specific people who were running this study. This sample of subjects is can be extremely biased. In a better study, there would be a larger sample size with more randomness in the recruiting process. Additionally, capturing more covariates would be a good idea in a future study so that randomness and bias can be more thoroughly evaluated with a covariate bias check. I'm totally fine w/ this as a limitation

The quality of step tracking was also a limitation in this study. The attrition experienced in this study may have animpact on the true results. It's not fully clear where attrition may have occurred, and in this study we were only able to make estimations. For example, if a fitness tracker or smart phone was dead for part of a day, there isn't a way for us to know this. In the data, we must assume that devices are never dead, and if steps were synced at all for that day, we assume it is complete for the entire day. It's possible that there's a correlation between people whose devices die frequently and their response to a step competition.

Finally, it may be worth blocking on cluster type at the beginning of the study and recruit large enough sample sizes accordingly. For example, it's possible that workplace competitions have different effects than competitions among friends or family. Blocking on these factors may help better understand if employee wellness programs are the right solution to help with health outcomes.

Really good job being honest in your reporting of the possible problems that might exist — I think — like you — that these are pretty trivial.

THERE ARE MORE NOTES ON THE LAST PAGE.

# Appendix A - Messaging to Volunteers

## 1. Initial Recruitment Message

*Hello!*

*As part of my Data Science Master's program at UC Berkeley, I'm taking a statistics course this semester. Two of my peers and I are studying people's activity and step habits for our course project.*

*We are recruiting volunteers to share their daily step count with us through the app Stridekick over the course of the 2 week experiment the last 2 weeks of March. Stridekick has cross fitness tracker platform compatibility!*

*If you choose to participate (which would be very kind of you), you will be entered into a lottery with the other participants for a chance to win 1 of multiple $25 Visa gift card. Plus, you'll be helping science and the advancement of humanity :wink:*

*To sign up, please complete this brief survey: https://goo.gl/forms/VSJBBSeDIfDOf8uV2*

*Thanks!*

*Annie Lane, Mursil Makhani, and Zach Merritt*

## 2. Stridekick Set up Qualtrics E-mail

*Thank you for signing up for our Steps Study as part of the UC Berkeley Master of Information and Data Science program. The step data collection starts in just a few days on Monday, March 18th, so please follow the directions below to get connected!*

*Here are the steps to connect to the study on Stridekick:*
*1. Get the Stridekick app on your phone*
*2. Create an account to connect your device*
*3. Friend our team @MIDS-Steps*
*4. Complete the confirmation survey at the link below*

***Follow this link to the Survey:***
*${l://SurveyLink?d=Take the Survey}*

*Or copy and paste the URL below into your internet browser:*
*${l://SurveyURL}*

*Check out our User Guide for a detailed step by step.*
*Find answers to your questions in our FAQ.*

*Thanks and happy stepping!*
*Annie, Mursil and Zach*

## 3. Control Group Saturday Qualtrics e-mail

*Happy weekend!*

*This is your friendly reminder to please* ***open the Stridekick app to automatically sync your steps*** *:)*

## 4. Treatment Group e-mail

*To add some fun to Week 2, we are inviting to you a 5 day leaderboard competition on Stridekick against other study participants, starting on Monday, March 25th.*

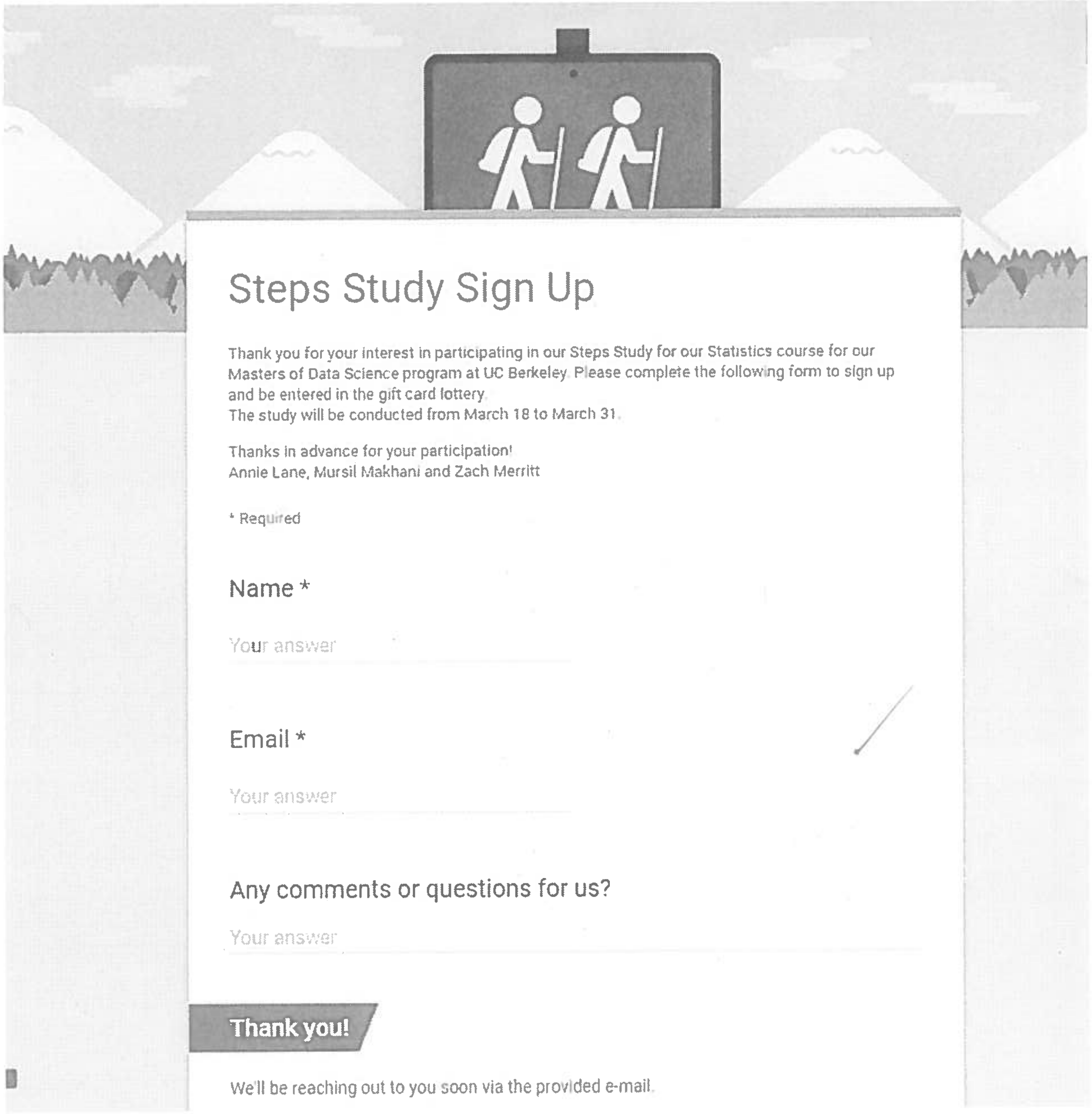
***The goal:*** *Get more total steps than your competitors during the 5 days. The first place finisher will receive a $35 Visa gift card. The second place finisher will receive a $15 Visa gift card.*

*You will receive an invite to the competition on Stridekick today - a notification will appear in the app and via an e-mail from Stridekick. Use either of these to* ***accept the invitation and officially join the competition****.*

*Please remember to refresh your steps on the app regularly to see where you stand!*

# Appendix B - Surveys

## 1. Initial Recruitment Survey



Steps Study Sign Up

Thank you for your interest in participating in our Steps Study for our Statistics course for our Masters of Data Science program at UC Berkeley. Please complete the following form to sign up and be entered in the gift card lottery.
The study will be conducted from March 18 to March 31.

Thanks in advance for your participation!
Annie Lane, Mursil Makhani and Zach Merritt

* Required

Name *

Your answer

Email *

Your answer

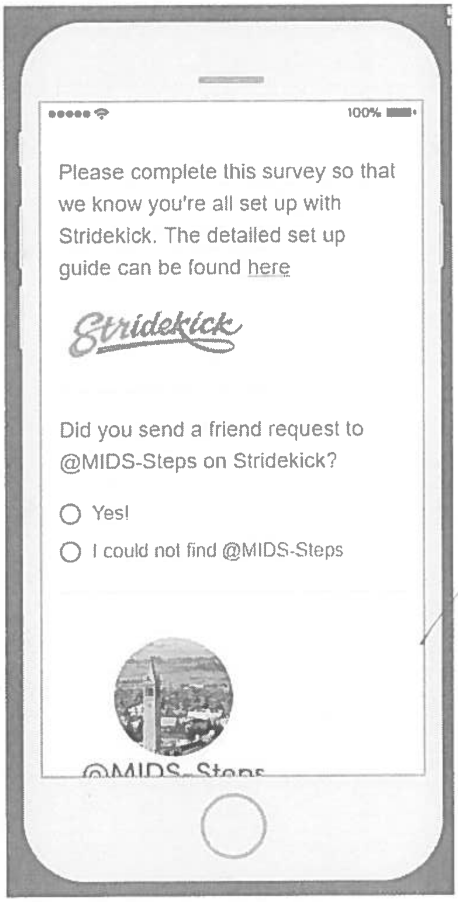
Any comments or questions for us?

Your answer

Thank you!

We'll be reaching out to you soon via the provided e-mail.

## 2. Stridekick Set up Survey



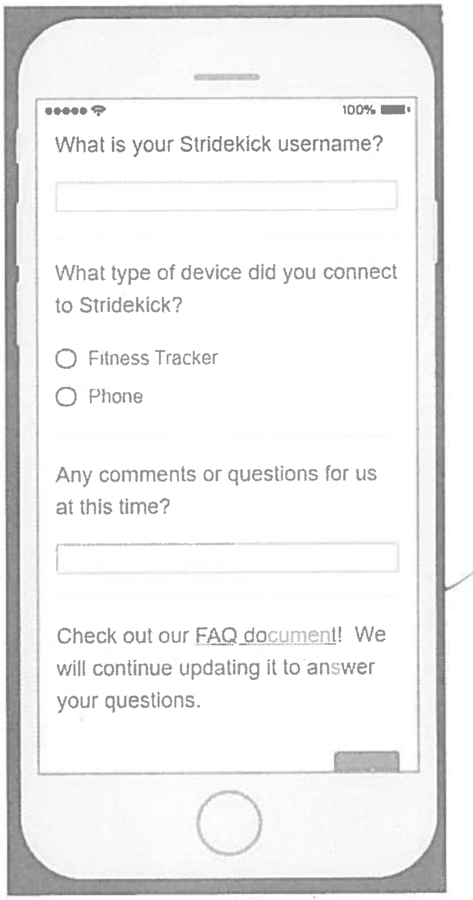
Please complete this survey so that we know you're all set up with Stridekick. The detailed set up guide can be found here

Did you send a friend request to @MIDS-Steps on Stridekick?

○ Yes!

○ I could not find @MIDS-Steps

@MIDS-Steps



What is your Stridekick username?

What type of device did you connect to Stridekick?

○ Fitness Tracker

○ Phone

Any comments or questions for us at this time?

Check out our FAQ document! We will continue updating it to answer your questions.

# Appendix C - Additional Participant Materials

## MIDS Step Study FAQ

*https://docs.google.com/document/d/1tbUF10CLqWoet6_cjPr__lwy62SP1pehe2K3e_aeK8w/edit?usp=sharing*

## Stridekick Set Up User Guide

*https://drive.google.com/file/d/1wP0lJlx4hqs1oxCuelEOcdjFMVt_282J/view?usp=sharing*

| CHART ID | TITLE EFFECT OF FITNESS COMPETITIONS | AUTHOR | DATE |
|---|---|---|---|
| DESCRIPTION | | | |

INPUT PROCESS OUT [redacted]

Team - this is a really well done job on this project. I think that you were really careful in the design and recruitment and data collection. This shows throughout. The paper is really well and clearly written, and the analysis is straightforward. This is a sign of a good design.

- I do wonder - however - if you have an exclusion restriction problem: Your treatment group receives ① Competition with others; but also, ② The chance to win $35 additional dollars. How can you be sure that it was the peer comp-? not the cash?