Blink, Buzz, Ring:

A Randomized Experiment on Smartphone Notifications and Stress

Cendy Lin, Andrew Lam, Dominic Delmolino, Matthew Nelson
UC Berkeley School of Information
W241 Experiments and Causal Inference, Section 1
April 2017

I. Introduction

For many people, technology pervades our everyday lives. The rise of the smartphone allows us to have the internet at our fingertips and many digital enterprises make the use of push notifications to increase engagement with their users. While these notifications may be useful, many of them may not be important. With an endless number of free apps vying for our attention, the smartphone can become a distraction from what we were previously focused on.

This paper seeks to address the following research question:
Will suppressing smartphone notifications reduce an individual's self-assessed stress levels?

Part of the impetus for this research question was driven by personal experience. The authors of this paper have observed the benefits of reducing notifications in their own lives. Whether this be focusing intently on a specific task, disconnecting from technology on vacation, or fully immersing yourself in an experience, there seems to be a compelling argument that we may need less screen time, not more. Given that many people own a smartphone and use it for both work and pleasure, the implications of this experiment are not insignificant.

To our knowledge, there have been few prior studies on this topic. Recently, there was "The Do Not Disturb Challenge" that occurred in 2016[1]; this, however, only examined 30 subjects and the treatment was only applied for 24 hours. This paper will expand on the existing literature with a study that has a larger sample size and examines the treatment over an extended period of time.

Our findings hint at supporting our hypothesis that turning off notifications reduce stress. However, we were met with quite a few challenges around attrition and noncompliance, which make it difficult to make substantive conclusions. In the next section, we will describe the experimental design and implementation. Thereafter, we will present our statistical analyses and main results, including discussions around attrition and noncompliance. Finally, we will conclude with our takeaways from the experiment and areas of future research.

II. Experimental Design Overview & Duration

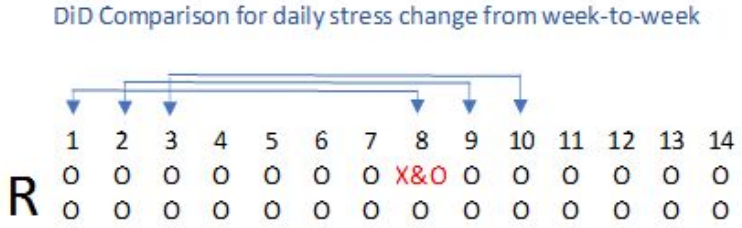
We sought to identify the impacts of phone notifications on stress levels by utilizing an experimental design where a group of control subjects were told explicitly to ensure their cell phone notifications remained on while a group of treated subjects were told to turn their notification settings off, and an assessment of daily stress levels would be recorded. The duration of the experiment was designed to be a two week study, with the first week used to gather

[1] https://pdfs.semanticscholar.org/1834/e42a4523e4b25933ca87a006df8160322aec.pdf

baseline data on individuals' stress levels with the following week measuring stress levels with treatment in place . There was concern that forcing subjects to keep their notifications off for too long (where it wasn't their choice to do so) would actually increase stress if they began to feel that they were not responding to important notifications in a timely manner. On the other hand, it was assumed stress levels would take at least a few days to stabilize to a new normal. It was decided that a one week period would allow subjects to experience any perceived benefits from the notification suppression, without initiating too heavy of a push-back that the instructions to remove notifications could be making their life more difficult (increasing the potential of non-compliance effects as we move away from the treatment start date).

Design Notation

In general, the research design looked like this, where the outcome variable was the difference in stress levels between week one and week two:



Estimand/Surveys

Our estimand, being the change in daily stress levels for each individual, was not able to be directly measured outside of utilizing calibrated heart rate monitors, blood pressure monitors, or other direct measures of physiological changes. Heart rate monitors have additional complications as it is almost impossible to tell whether heart rate increased due to stress, physical exertion, or some other factor and any other method for directly recording stress would require in-person contact with all subjects which was not feasible given our resources. It was therefore decided to utilize self-assessment surveys that would be filled out by each subject at the closing of each day.

These surveys would ask a number of questions relating to general health (such as the quality/quantity of sleep or their general levels of happiness today) in order to distract from the question directly relating to our estimand: "What was your stress level today?". Stress levels were recorded on a likert scale of 1 to 5, with 1 being "non-existent" and 5 being "extremely high".

The outcome variable of interest would be the change in stress level for every day of the week, for example, the difference in stress between the baseline week Monday and treatment week

Monday. There would therefore be seven change-in-stress observations per individual, assuming they completed all surveys. In addition, the surveys would ask a question regarding compliance to roughly determine if each subject remained in treatment (kept their notifications off) throughout the duration of the day. In order to encourage participants to complete these daily surveys, they were placed in a draw for five $100 Amazon gift cards, with individuals who completed at least 90% of the surveys entered once into the raffle, and individuals who completed 100% of the surveys entered twice. Honesty in the surveys was something that could not be enticed, though a comment at the top of the surveys was included which encouraged participants to answer the questions honestly and that there would be no penalty (in terms of entry to the gift card draws) for individuals who failed to stay in treatment (and therefore there was no reason to lie about it).

The self-assessment surveys were to be sent out every evening at the same time with a request for it to be filled out within a 24 hour period (a sample survey is included in the Appendix). Prior studies have shown that self-assessments are much more accurate when completed immediately after an activity when that activity (or day's activities, in this case) is still fresh in the mind. We did find a few individuals who completed multiple surveys in one day, which would be indicative of participants 'playing catch up'. These data points were kept in the analysis and we assumed individuals responded to the daily surveys chronologically.

As daily stress levels are highly variable and can vary significantly due to external events, a proper randomization procedure was paramount to the success of this approach. In addition, stress levels tend to vary on different days of the week (are likely lower on the weekends) and therefore the surveys would need to be completed daily but distributed in weekly increments. The high variability of stress, combined with a likert scale of only 1-5 in the daily stress survey, provides some potential for the stress results to overlap and appear imprecise. Power calculations were completed to determine the number of subjects needed in order to achieve an 80% likelihood of finding a significant result which resulted in a requirement to have a total population of 126 participants completing the study, with 63 in treatment and 63 in control to significantly detect a difference of 0.5 on the likert scale (assuming that since most people tend toward the middle on a likert scale for our control group, and that results for the treatment group may vary +/- 1 point). To detect a difference of 0.75 on the likert scale, we needed a population of 56 participants. Since we had a smaller sample size (38 and 39 people in control and treatment respectively), we decided to use individual change-in-stress observations, effectively giving us a sample seven times our original sample size (seven change-in-stress observations per individual), thereby providing more power in our experiment.

The main risk to this approach in that failures to respond to either a baseline or treatment week survey would result in missing (attrited) change-in-stress observations, and the resulting average

treatment effect could be biased towards the stress levels of the individuals who submit all surveys (and who may be fundamentally different, with a different potential outcome to treatment, than those individuals who do not submit all the surveys). The Amazon gift card enticement should help increase daily survey participation, though there would likely be individuals who choose that the reward is not worth their effort and fail to respond to all of the surveys. Additionally, some subjects may decide to respond to the surveys only when they feel good or only when they feel bad, incorporating bias into the results as well. Compliance in both adherence to treatment and response to the daily surveys is paramount to the success of this approach.

Exclusion / Non-Interference Assumptions

All control and treatment participants were required to turn on their blue-light suppression settings (to distract from the true purpose of the experiment), and both control and treatment groups received instructions on how to set their Do Not Disturb settings. Additionally, the only interaction between subjects and experimenters was standardized emails so that there would not be any additional risk to the excludability assumption.

Within this experimental design there was some risk of the non-interference assumption being violated (spillover between treatment and control groups), as a large number of the subjects would be recruited from family and friends. Especially in the case of family living under the same roof there would likely be some discussion between subjects regarding their instructions after the baseline week, especially if their instructions differ as a result of being placed separately into treatment and control. While this feature could have been blocked on (same address) it was not one of the questions we asked in the application and therefore blocking on this covariate did not occur.

Treatment

Taking into considerations that every phone user will likely have a different assortment of apps pushing notifications (to their screen, to vibrate or to an auditory alert) it was decided that the treatment groups would be told to utilize their Do Not Disturb settings which naturally affect all notifications regardless of app style (email, social media, news, etc.) and could be applied across the treatment group without substantial individual instruction. In general, the treatment was specified to be: Turn ON Do Not Disturb settings (excluding phone calls) which would suppress lights, vibration, and sound. It was decided that notification banners seen upon opening the screen would be fine, with the idea being that if an individual is deep in work they should not be pulled out of it by a 'forced notification'. If they choose to click the home button on the phone and turn on the screen, they can look at the new contents of any app and that would be on their own initiative. The treatment was designed to address the concept of removing distractions, which individuals cannot willfully prevent at any given moment, while retaining the primary use

of the phone (to receive phone calls). Prior studies had shown the propensity of fully blocking incoming phone calls as a stress inducer to people who were afraid that they would miss important messages from friends and family[2].

In order to help distract from the exact nature of this study, it was decided that both the control and treatment groups would be asked to complete an additional modification to their smartphone: turn on blue-light suppression settings during the treatment week of the study. If the design of the experiment excluded this step, it would be likely that the control group could infer that they were part of the control subset of participants and provide biased results in their surveys, or fail to respond to the surveys in general. By asking both groups to turn on blue-light suppression settings (through Night Shift on iOS or Twilight on Android) there was a willful attempt to ensure active participation by the compliers in the control group. The selection of blue-light suppression as a distraction was done because it was hypothesized that it would not have a significant effect on stress levels, and if it did its effects would at least be balanced between the control group compliers and treatment group compliers.

The study was administered in the following way: both treatment and control groups were sent an introductory email at the beginning of the study (day 1), thanking them for their participation and explaining that they would be sent a daily survey every evening for the next 14 days. They were instructed to add the email sender (Matthew Nelson) to their contact list to avoid any emails from entering the junk mail folder, which we discovered was an issue during our pilot study. A single email was sent to each participant every evening in the first week containing a link directing them to a short Google Forms survey. This email and link did not change throughout the week. On the 7th day, a follow-up email with detailed treatment instructions was sent to the control group, indicating that they should turn on their blue-light suppression settings and ensure that their Do Not Disturb settings were set to **OFF**. The same email was sent to the treatment group with the exception that they were asked to turn their Do Not Disturb settings **ON**. This was explicitly asked because, as addressed in the power calculations, there was some concern that any effect from the change in notification settings on stress levels would be lost in the natural variation of self-reported stress. In order to evaluate the *maximum plausible effect* we would need to ensure that individuals in the control group have their notifications on and then compare them to the treatment group who is explicitly told to turn their notifications off.

[2] https://pdfs.semanticscholar.org/1834/e42a4523e4b25933ca87a006df8160322aec.pdf

Week	1							2						
	Mon	Tues	Wed	Thurs	Fri	Sat	Sun	Mon	Tues	Wed	Thurs	Fri	Sat	Sun
Assigned To:	4/3	4/4	4/5	4/6	4/7	4/8	4/9	4/10	4/11	4/12	4/13	4/14	4/15	4/16
Control	DS3	DS3	DS3	DS3	DS3	DS3	DS3	DS2	DS2	DS2	DS2	DS2	DS2	DS2
Control	DS3	DS3	DS3	DS3	DS3	DS3	DS3	DS2	DS2	DS2	DS2	DS2	DS2	DS2
Control	DS3	DS3	DS3	DS3	DS3	DS3	DS3	DS2	DS2	DS2	DS2	DS2	DS2	DS2
Treatment	DS3	DS3	DS3	DS3	DS3	DS3	DS3	DS1	DS1	DS1	DS1	DS1	DS1	DS1
Treatment	DS3	DS3	DS3	DS3	DS3	DS3	DS3	DS1	DS1	DS1	DS1	DS1	DS1	DS1
Treatment	DS3	DS3	DS3	DS3	DS3	DS3	DS3	DS1	DS1	DS1	DS1	DS1	DS1	DS1

DS3 = Daily Survey3 (Baseline Data Gathering)
DS1 = Daily Survey 1 (Treatment)
DS2 = Daily Survey 2 (Control)
Treatment Instructions
Control Instructions
Start of Experiment Instructions

Subjects

Subjects were recruited primarily through the experimenters' direct social networks, specifically Facebook and Twitter. An invitation to participate was posted on Facebook and was also run through two separate Twitter promotions. Subjects were invited to click on a link which took them to an application form in Google Forms. A number of covariate gathering questions were asked of the subjects, specifically regarding their daily smartphone habits, age, gender, and phone operating system. In total, **77 subjects** responded in time and were selected to be included in the study. 6 of the first responders were selected to be a part of the pilot study, which was run three weeks prior to the start of the main experiment.

Randomization

Subjects were blocked on **gender** (M/F), **age** (<24/25-54/55+), **phone operating system** (Android / iOS) and whether or not they had **one phone for work and personal use** (a proxy for high/low frequency work notifications). Subjects were then randomly assigned to treatment and control within these blocks, ensuring that the total number of control and treatment subjects was approximately equal (38 and 39 respectively). A quick check of the control/treatment

balance within blocks showed that blocks were balanced:

Block	Control	Treatment	Total
1	1	1	2
3	7	7	14
4	6	7	13
5	3	3	6
6	2	2	4
10	4	3	7
11	0	1	1
12	1	0	1
16	1	0	1
21	3	3	6
22	2	3	5
23	3	3	6
24	0	1	1
28	1	1	2
30	1	1	2
Pilot	3	3	6
Totals	**38**	**39**	**77**

Covariate Balance

We examined the covariate balance between treatment and control groups to ensure we were comparing apples-to-apples. Our balance check on our initial group of 77 participants is below:

	Difference in Means		
Covariate	**Control Mean 38 people**	**Treatment Mean 39 people**	**P-value of Difference**
Male	0.47	0.51	0.735
Age (<44)	0.84	0.82	0.803
Apple Phone	0.79	0.85	0.526
Work+Personal phone (vs. personal phone only)	0.26	0.31	0.670
Phone Use (at least once per hour)	0.95	0.92	0.670
Relationship (Family/Friend vs. Other)	0.55	0.62	0.582

Attrition

The main challenge with the experiment design relates to attrition. It was unlikely that an experimental design utilizing daily surveys submitted by incentivized subjects would be successful in receiving 100% of the requested surveys. While the submission of the daily survey requires only a small commitment, it is highly likely that subjects will miss/skip days or decide that they no longer want to participate in the study altogether because the potential reward does not match the effort. This lack of completion surrounding the recovery of the main outcome variable causes problems during the evaluation stage. Individuals who choose to return 100% of their surveys may be fundamentally different than those individuals who only return 50% and this prevents the use of an overall post treatment average (taking all post-treatment observations in the treatment group and dividing them by the number of responses) because those subjects with more surveys submitted will bias the overall average towards their recorded stress levels. Another challenge of attrition with this design is in regards to those who submit less than the full number of requested surveys, and their motivation for doing so. If they chose to submit a survey only when they were feeling good, or when they actually followed through on the treatment, this would introduce bias into the estimate of the average treatment effect. The noise that non-responses added to the analysis was of main concern as the main work-around for attrition includes analyzing the banded range of responses, so the more attrition, the more variability.

In order to address these attrition challenges the researchers decided to:

1) Restrict the subject pool to individuals who provided at least one baseline week and one treatment week survey
2) Use a conservative banded range analysis and fill in all missing daily survey data with the most extreme potential responses (1 and a 5 for stress levels) on the restricted subject pool

Restricting the Subject Pool

Since we have a number of individuals who provided little to no survey responses, we restrict the analysis to the group of participants who have at least one baseline week and one treatment week survey responses. This is because we assume that individuals who had no survey responses either never received our emails (e.g. did not follow instruction to add us to their contact list) or "opted-out" of the experiment before it even began. We also had individuals who had no treatment week responses, who we assume also dropped out after the baseline week for unknown reasons. It should be noted that all 77 original participants continued to receive baseline and treatment week surveys through the whole experiment, but we decide to restrict our subject pool in order to reduce the noise that would be introduced by these individuals.

We had the following type and count of attritors:

1. No responses to baseline and treatment week surveys: **15**
2. Some responses to baseline week surveys but no responses to treatment week surveys: **6**
   - **Total "full" attritors: 21 of 77** (27% of total)

With these individuals removed, we now have **25** individuals in the control group and **31** individuals in the treatment group. We acknowledge that by restricting our subjects, we are throwing out some data collected by individuals who provided some baseline week responses but no treatment week responses. Between the 21 "full" attritors, there were a total of 24 survey responses (where we should have had 21 x 14 or 294 surveys). We also note that amongst the remaining group, there is still missing data, but we use the restricted group in order to exclude noise that would be introduced from the 21 "full" attritors who provided little to no information.

To ensure that our randomization still works and that we are still comparing apples-to-apples, we conduct a covariate test on the restricted set:

	Difference in Means		
Covariate	**Control Mean 25 people**	**Treatment Mean 31 people**	**P-value of Difference**
Male	0.40	0.52	0.39
Age (<44)	0.84	0.81	0.75
Apple Phone	0.80	0.90	0.30
Work+Personal phone (vs. personal phone only)	0.24	0.29	0.68
Phone Use (at least once per hour)	0.92	0.90	0.77
Relationship (Family/Friend vs. Other)	0.72	0.71	0.93

We also conducted an F-test checking for randomness of our restricted group against the prognostic covariates and did not get a significant result, demonstrating that the restricted control and treatment group still appear random:

```
Analysis of Deviance Table

Model 1: treat_code ~ 1
Model 2: treat_code ~ 1 + male + age + relationship + apple + personal_phone +
    phone_use
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        55     76.988
2        49     72.731  6   4.2577   0.6418
```

We also examined the correlation between these 21 attrited individuals and prognostic covariates across both treatment groups:

	Correlation with Attrition		
Covariate	**Control 38 people**	**Treatment 39 people**	**Total 77 people**
Number "Full" Attritors	**13**	**8**	**21**
Treatment Group	NA	NA	-0.154
Male	0.205	-0.013	0.095
Age (<44)	0.008	0.072	0.042
Apple Phone	-0.036	-0.311	-0.165
Work+Personal phone (vs. personal phone only)	0.073	0.074	0.065
Phone Use (at least once per hour)	0.170	0.147	0.161
Relationship (Family/Friend vs. Other)	-0.467	-0.382	-0.430

We find that most correlations are weak with a few exceptions:

- "Relationship" is negatively correlated with attrition across both groups (i.e. individuals who are either friends or family of the researchers appear less likely to be "full" attritors)
- In the control group, men are more likely to be "full" attritors
- In the treatment group, individuals with Apple phones are less likely to be "full" attritors

However, given that our randomization check (using the F-test) and our covariate balance check on the restricted subject pool checks out, and given the amount of noise added by including these "full" attritors in the analysis, the researchers decided to conduct the analysis on the restricted subject pool of 56 individuals.

Within the restricted subject pool, we still had a number of attrited observations (where baseline week information is missing, treatment week information is missing, or both). We also looked at the relationship between the attrited observations (not individuals) amongst the restricted sample and their covariates. A summary table is provided below:

	Control 25 people	**Treatment 31 people**	**Total 56 people**
Total Expected Observations	175	217	392
Number Missing Observations (% Missing or Attrited)	38 (22%)	61 (28%)	99 (25%)
Correlation Between Attrited Change-in-Stress Observations and Covariates			
Treatment Group	NA	NA	0.073
Male	0.023	0.134	0.095
Age (<44)	0.041	0.047	0.041
Apple Phone	0.125	0.031	0.086
Work+Personal phone (vs. personal phone only)	-0.166	-0.152	-0.153
Phone Use (at least once per hour)	-0.151	0.135	0.017
Relationship (Family/Friend vs. Other)	0.020	-0.097	-0.048
Weekday	0.083	0.015	0.044

With the restricted group, we have approximately 25% missing observations overall, with the treatment group having a slightly higher missing proportion (28% vs 22% in control). As an aside, if the "full" attritors were to be included, the percentage of attrited responses would be closer to 50% since the "full" attritors, which made up 27% of the original sample, contributed over 20% of the overall attrited responses. When looking at the correlation between covariates and attrited responses in the restricted sample, it appears that most relationships between covariates and attrition are weak. In the analysis, these attrited responses would be filled in with the minimum and maximum stress levels to get a banded ATE range.

Noncompliance

Noncompliance is a significant issue within this experimental design, specifically because the experimenters are blind to whether or not a subject followed through with treatment (unable to personally check their phone settings periodically). Compliance levels were assessed by asking subjects to self-declare whether or not they had followed through with treatment within the 24 hours prior to the daily survey response (i.e. "Did you leave Do Not Disturb settings ON?"). Clearly, there is no incentive for subjects to answer this question truthfully, and there may in fact be a proclivity for subjects to declare themselves as compliers even in the case that they did not comply fully that day because they are behaving as they believe they are expected to behave. Since the compliance question asked of treatment group individuals is the best indicator of compliance considering the restrictions, the researchers will use this as a proxy. A summary of non-compliance is provided below:

Total Number Non-Missing Treatment Observations	176
Number Non-Compliance (% of Non-Missing Observations)	51 (29%)
Alpha (% complied)	71%

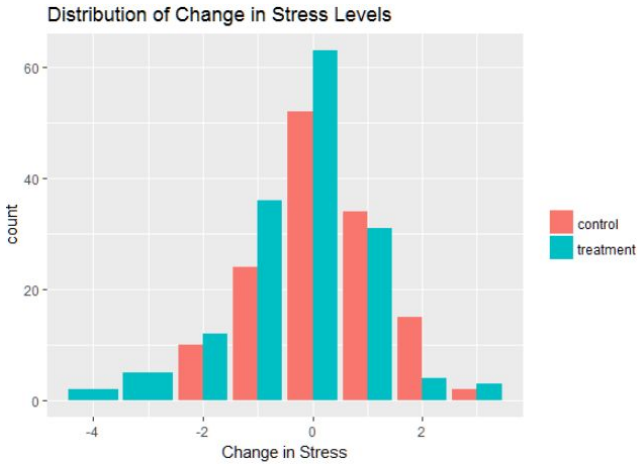
Since we are interested in the effect for individuals who actually take treatment, we are most interested in the Complier Average Causal Effect (CACE), so we will divide our ATE by 71% to get the CACE.

III. Main Results

Analysis Ignoring Missing Observations

Our initial analysis was conducted on the restricted subject pool, ignoring the missing observations.

When looking at the initial distribution of change-in-stress, we see that generally, the treatment group is seeing more of a decrease in stress than the control group. It also appears that the control group generally experiences an increase in stress between baseline and treatment weeks, which seems promising.



We decided to use a linear regression as our model, with change-in-stress as the main

outcome variable, treatment (0 if control or 1 if treatment) being our main independent variable. This is based on a simplifying assumption that stress on the likert scale is linear, which it may not be because self-reports on stress are subjective. We also control using the 'weekday' covariate to minimize variability introduced by the day of the week. Our model is thus:

$$\Delta Stress = b_0 + b_1 treatment + b_i weekday_i$$

We use clustered standard errors clustering at the individual level since individuals will have similar potential outcomes. Based on our initial analysis ignoring attrition, our results are:

```
t test of coefficients:

                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.019630   0.187643  0.1046 0.916758
treattreatment   -0.416472   0.153984 -2.7046 0.007249 **
weekdayMonday     0.631425   0.232569  2.7150 0.007032 **
weekdaySaturday  -0.055725   0.186375 -0.2990 0.765164
weekdaySunday    -0.190570   0.219047 -0.8700 0.385036
weekdayThursday   0.256699   0.218350  1.1756 0.240724
weekdayTuesday    0.558045   0.282203  1.9775 0.048954 *
weekdayWednesday -0.059405   0.294513 -0.2017 0.840292
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Initial results look promising, with an estimated treatment effect of **-0.42**, which is significant at the 5% level. Dividing by our alpha (or percentage of compliant observations in the treatment group), we get a CACE of **-0.58**, demonstrating that amongst non-attrited responses, we are seeing about a half point drop in stress when comparing the treatment group to the control group.

However, this analysis does not include the 25% of missing observations and may therefore be biased. We therefore decide to run a banded range analysis by filling in missing stress values with 1's or 5's to get the maximum and minimum treatment effect (+4 maximum increase in stress or -4 maximum decrease in stress).

Dealing with Attrition: Banded Range

Maximum Effect

In order to calculate the maximum effect, we want to fill in the control group with the maximum **increase** in stress levels, so missing control observations would have 1 inserted in the baseline week and 5 inserted in the treatment week. We want to fill in the treatment group with the maximum **decrease** in stress levels, so missing

treatment observations would have 5 inserted in the baseline week and 1 inserted in the treatment week. Doing this, our distribution of change-in-stress observations looks like the graph shown on the previous page, where we can clearly see where the missing control and treatment values were inserted.

Our linear regression on our data set with maximum effect values filled in and clustered standard errors per individual is as follows:

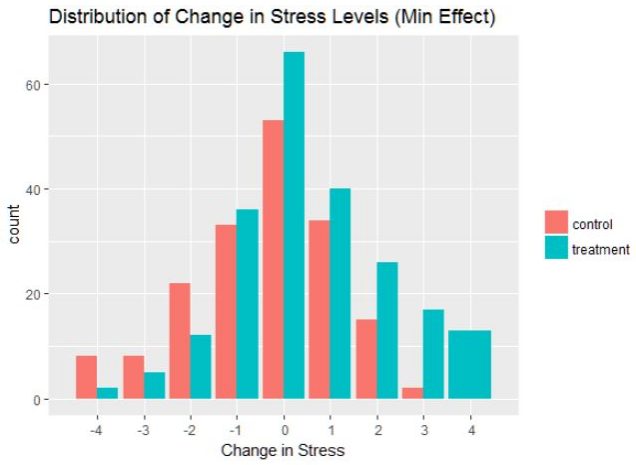
```
t test of coefficients:

                 Estimate Std. Error t value  Pr(>|t|)
(Intercept)      0.693980   0.253856  2.7338  0.006551 **
treattreatment  -1.447189   0.210603 -6.8717 2.574e-11 ***
weekdayMonday    0.392857   0.297299  1.3214  0.187148
weekdaySaturday -0.071429   0.200594 -0.3561  0.721972
weekdaySunday   -0.285714   0.260602 -1.0964  0.273607
weekdayThursday  0.017857   0.233542  0.0765  0.939091
weekdayTuesday   0.178571   0.275876  0.6473  0.517831
weekdayWednesday -0.250000   0.267532 -0.9345  0.350651
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Filling in values with the maximum effect gives us a greater decrease in stress levels, with an ATE of **-1.45** Our CACE would thus be **-2.04**. We expect to see a larger decrease in stress when compared with our model that excludes missing values since we are essentially filling in "NAs" with negative values. It is also interesting to see how stress changes by day of the week (when compared to Friday): Monday seems to experience an increase in stress, whereas weekend days experience a decrease in stress, although none of these coefficients are significant.

Minimum Effect

To fill in missing values with the minimum treatment effect, we want the missing control values to have the maximum **decrease** in stress, and the missing treatment values to have the maximum **increase** in stress. Therefore, missing control group values will receive a 5 for the baseline week and a 1 for the treatment week. Missing treatment group values will receive a 1 for the baseline week and a 5 for the treatment week. Our new distribution with minimum treatment effect filled in for missing values is shown at right, where one can see that the treatment group now has many more +4 observations, and the control group has many more -4 observations.



Running our linear model against the data with minimum effect values filled in and clustered at the individual level, we get the following result:

```
t test of coefficients:

                 Estimate Std. Error t value  Pr(>|t|)
(Intercept)      -0.44929    0.24105 -1.8638 0.0631062 .
treattreatment    0.87613    0.24783  3.5352 0.0004571 ***
weekdayMonday     0.32143    0.27588  1.1651 0.2446934
weekdaySaturday  -0.28571    0.23386 -1.2217 0.2225549
weekdaySunday    -0.28571    0.28017 -1.0198 0.3084620
weekdayThursday   0.37500    0.25005  1.4997 0.1345093
weekdayTuesday    0.39286    0.31248  1.2572 0.2094358
weekdayWednesday  0.10714    0.30171  0.3551 0.7226992
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Unfortunately, when filling in missing values with the minimum effect, our average treatment effect now appears positive, **0.88** increase in stress and is significant at the 5% level. Our CACE under this scenario is **1.23.**

The following table summarizes our three models, 1) ignoring attrition, 2) using maximum effect for missing values, and 3) using minimum effect for missing values:

	Dependent variable:		
	Ignore Attrition	Max Effect	Min Effect
	(1)	(2)	(3)
treattreatment	-0.416***	-1.447***	0.876***
	(0.154)	(0.211)	(0.248)
weekdayMonday	0.631***	0.393	0.321
	(0.233)	(0.297)	(0.276)
weekdaySaturday	-0.056	-0.071	-0.286
	(0.186)	(0.201)	(0.234)
weekdaySunday	-0.191	-0.286	-0.286
	(0.219)	(0.261)	(0.280)
weekdayThursday	0.257	0.018	0.375
	(0.218)	(0.234)	(0.250)
weekdayTuesday	0.558**	0.179	0.393
	(0.282)	(0.276)	(0.312)
weekdayWednesday	-0.059	-0.250	0.107
	(0.295)	(0.268)	(0.302)
Constant	0.020	0.694***	-0.449*
	(0.188)	(0.254)	(0.241)
Note:			*p<0.1; **p<0.05; ***p<0.01

Our banded range ATE thus ranges from **-1.447 to 0.876**, both values are significant when comparing against the null hypothesis that there is no treatment effect.

Our CACE (where alpha = 0.71) further widens the gap, ranging from **-2.04 to 1.23.**

As an aside, including the “full” attritors in the analysis would have given us a banded range of -3.08 to 2.64, or a CACE of -4.59 to 3.94, so removing these individuals did in fact help to reduce noise. Unfortunately, since our range(s) include 0 (no difference in stress levels), we cannot ultimately conclude that turning off notifications leads to a reduction in stress.

IV. Conclusion

The use of a smartphone's "Do Not Disturb" functionality to suppress application notifications may have a positive impact on an individual's assessment of their stress level, but gaps in reported data required conservative analysis that ultimately made such a conclusion difficult to obtain.

Our study had several areas of strength including sampling a larger number of participants than previous studies, strong blocked randomization with balanced covariates across treatment, control and attrited groups, longer periods of stress measurement than previous studies and the concept of stress related to day-of-week feelings by participants. However, even though we performed a pilot study to flush out errors, our completed full study suffered from several limitations.

Since we relied on our participant's understanding of instructions and their honesty with regard to both initially following them and continuing to comply with directed smartphone settings, we exposed our results to variations in participant comprehension, execution expertise and honesty with regard to compliance. Using self-reported and self-assessment survey tools meant that we were unable to secure timely data due to subjects attention (and inattention) to daily survey completion, while also being constrained by the subjects ability to accurately self-assess their stress levels. Subjects may be exhibiting an enhancement bias on their survey results by reporting less stress, or avoiding responses when they are experiencing high stress.

With a promising initial result minimized by the conservative data analysis necessitated by survey and assessment issues, follow up research should be undertaken to address study shortcomings.

In the ideal follow-up study, one would always like to obtain a larger pool of subjects in order to increase the likelihood of detecting small changes in stress levels. We had reasonably decent success through the use of paid, online advertising directing interested subjects to registration pages and surveys. Further, limited investigation led to online sites which can help identify research participants. Increasing the number of subjects from 75 to about 300 would have helped us better detect our estimated treatment effect. Since subject acquisition costs seemed to be about $5 / subject, moving to this larger subject pool can be done rather effectively. Mobile device management applications (such as AirWatch, from VMWare) could be used to control subject's Do Not Disturb settings[3], provided subjects agree to the continued installation of the software. Doing so would consistently set the treatment, and allow accurate measures of compliance. Our

[3] https://my.air-watch.com/help/9.1/en/Content/Expert_Guides/System_Settings/DevicesUsers_General_Privacy.htm

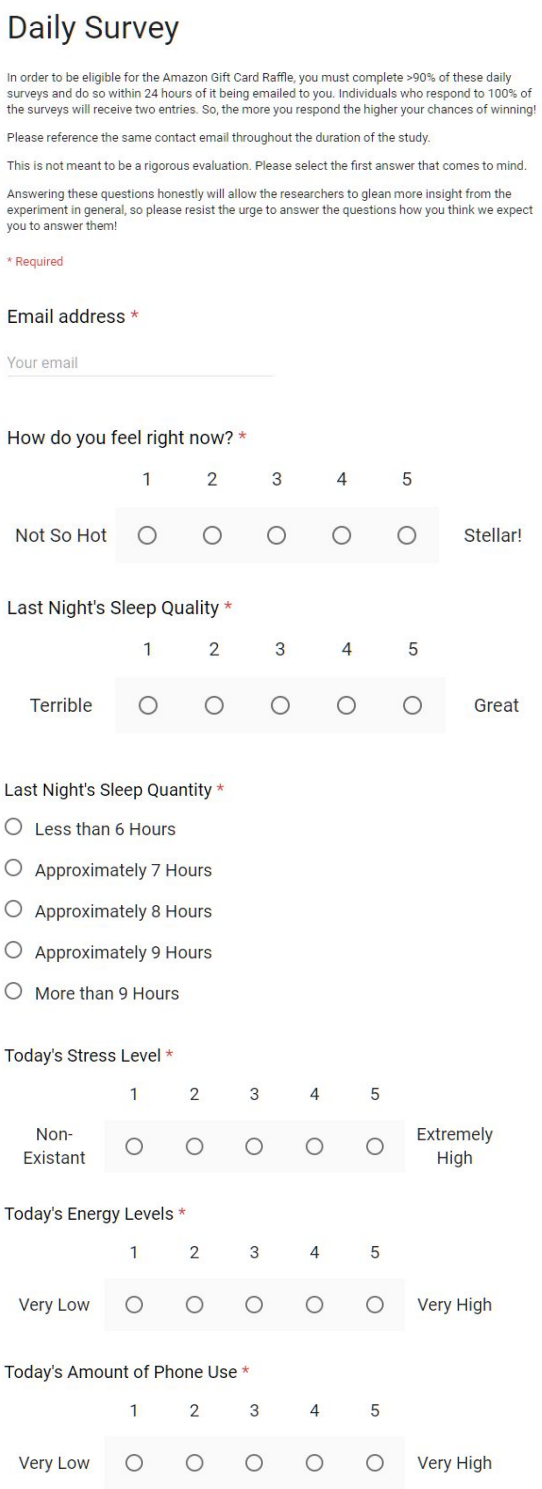
survey instrumentation suffered through the lack of unique assignments of survey result identification, including the use of re-typed email addresses with each survey being used to link results to subjects (leading to the requirement to clean email address entries for matching and linkage to subjects) and the lack of a survey ID corresponding to the day a survey applied to. Using a combination of a unique, system-generated, immutable subject id included with a survey day id in the actual survey link and response would have addressed this challenge.
Stress measurement and assessment could be enhanced through a variety of means, with the easiest using the 1983 10-point Perceived Stress Scale (PSS)[4] (the challenge being that the PPS assessment is a monthly survey, which while demonstrating reliable results, has limit history applied to daily assessments). However, the PPS suffers from all self-reported surveys related to self-enhancement bias and honesty. A more consistent person-to-person measurement approach based on physiology may be more useful. There are several "stress measurement" apps in mobile app stores which use smartphones to measure heart rates during controlled breathing sessions as a proxy for stress. While the correlation between those physical measurements and actual (or, more to the point) perceptions of stress is unclear, they do provide a more consistent measurement item that isn't subject to honesty and enhancement bias. Finally, attrition and missing results could be addressed through the use of follow-up surveys to gather information about why subjects may have left the study or failed to respond to surveys as a way to design better covariates and to determine better banded ranges / extremes to use for imputing missing data.

With the continued spread of smartphones (with ever larger memory allowing for more and more apps to be installed with their attendant notifications) and their potential to add stress to our lives, options for this study to be repeated seem to be of value. With relatively minor changes, it should be possible to determine if our initial results can be borne out with more reliable data.

[4] Cohen, Sheldon, et al. "A Global Measure of Perceived Stress." *Journal of Health and Social Behavior*, vol. 24, no. 4, 1983, pp. 385–396., www.jstor.org/stable/2136404.

V. Appendix

Daily Survey - Treatment Week Surveys for Treatment Week



Daily Survey

In order to be eligible for the Amazon Gift Card Raffle, you must complete >90% of these daily surveys and do so within 24 hours of it being emailed to you. Individuals who respond to 100% of the surveys will receive two entries. So, the more you respond the higher your chances of winning!

Please reference the same contact email throughout the duration of the study.

This is not meant to be a rigorous evaluation. Please select the first answer that comes to mind.

Answering these questions honestly will allow the researchers to glean more insight from the experiment in general, so please resist the urge to answer the questions how you think we expect you to answer them!

* Required

Email address *

Your email

How do you feel right now? *

	1	2	3	4	5	
Not So Hot	○	○	○	○	○	Stellar!

Last Night's Sleep Quality *

	1	2	3	4	5	
Terrible	○	○	○	○	○	Great

Last Night's Sleep Quantity *

- ○ Less than 6 Hours
- ○ Approximately 7 Hours
- ○ Approximately 8 Hours
- ○ Approximately 9 Hours
- ○ More than 9 Hours

Today's Stress Level *

	1	2	3	4	5	
Non-Existant	○	○	○	○	○	Extremely High

Today's Energy Levels *

	1	2	3	4	5	
Very Low	○	○	○	○	○	Very High

Today's Amount of Phone Use *

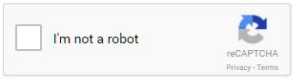
	1	2	3	4	5	
Very Low	○	○	○	○	○	Very High

Did you leave your blue light suppression settings (Twilight / Night Shift) ON all day? *

- ○ Yes
- ○ No

Did you leave your Do Not Disturb settings ON all day? *

- ○ Yes
- ○ No



SUBMIT